REDAÇÃO E MEDIDA DA EXPRESSÃO ESCRITA: ALGUMAS CONTRIBUIÇÕES DA PESQUISA EDUCACIONAL

HERALDO MARELIM VIANNA *

RESUMO

O artigo apresenta uma revisão de estudos empíricos realizados sobre a fidedignidade e a validade das provas de dissertação utilizada como instrumento de medida da capacidade de expressão escrita.

SUMMARY

Essay test dissertation and measurement of writing ability: contributions of educational research — This paper contains a review of empirical studies on reliability and validity of dissertation employed in measurement of writing ability.

A literatura brasileira sobre os instrumentos aplicados nos Concursos Vestibulares, para acesso ao ensino superior no Brasil, ainda é bem reduzida. Alguns poucos estudos empíricos sobre o assunto começam, entretanto, a ser divulgados, destacando-se, entre outros, os relativos à ponderação de provas (BARROSO, 1972), à fidedignidade de provas objetivas (BREEN, 1975), à análise de itens objetivos (RODRIGUES, 1972), à fidedignidade da Taxonomia de Bloom (SANCHEZ, 1972), aos testes de aptidão (RODRIGUES, 1974), à análise estatística de provas (MORAES e ANDRADE), 1970), às características psicológicas de vestibulandos (MORAES e ANDRADE, 1971), e os estudos de validade preditiva de provas objetivas na seleção de candidatos a escola de medicina veterinária (BARROSO *et alii*, 1972) e a escolas de medicina humana (BARROSO, 1972). Os dados coletados por intermédio dos Concursos Vestibulares permitiriam a elaboração de um amplo programa de pesquisas, conforme destaque de BESSA (1975), entretanto, no que se refere à análise do instrumental, pouco tem sido feito. A redação, em particular, ora introduzida como componente da prova de Comunicação e Expressão, em alguns Concursos Vestibulares, ainda não foi objeto de pesquisas empíricas; assim sendo, serão consideradas aqui algumas contribuições da literatura estrangeira.

Problemas relativos à fidedignidade

As pesquisas de STARCH e ELLIOTT (1912, 1913a e 1913b), realizadas há mais de sessenta anos, são consideradas clássicas no campo da docimologia educacional. As investigações procuraram determinar a amplitude de variação e a fidedignidade de notas atribuídas por professores a provas de dissertação em Inglês, Matemática e História. Os resultados gerais dessas pesquisas mostraram que: 1.º — existe uma grande amplitude de variação entre as notas dos professores e, em muitos casos, essa amplitude, para uma única prova, foi de 35 a 40

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

pontos; 2.º — a variabilidade ou não-fidedignidade das notas foi tão grande em uma área de conhecimento quanto em outras, positivando os pesquisadores que a variabilidade não é função do assunto, mas sim do examinador e do método de exame. Finalmente, a partir dos dados coletados, STARCH e ELLIOTT concluíram que a grande variabilidade das notas, para um mesmo trabalho de dissertação, tende a desacreditar a justiça e a precisão desse método de avaliação escolar.

Ao contrário de STARCH e ELLIOTT, que pesquisaram na área do ensino de segundo ciclo *(high school)*, ASHBURN (1938) realizou trabalho empírico sobre a correção de dissertações em nível universitário. Inicialmente, o experimento baseou-se em dissertações elaboradas por 65 estudantes de Literatura; posteriormente, a mesma pesquisa foi repetida com 75 estudantes e as dissertações versaram sobre Literatura e História. Em ambos os casos, os trabalhos foram corrigidos independentemente por grupos de três professores especialistas na matéria. Ao elaborarem o relatório final, os professores de Literatura revelaram não haver unanimidade no grupo sobre o que fosse uma boa resposta em questão de dissertação. O grupo de História, por sua vez, declarou estar convencido de que não poderia corrigir as dissertações com qualquer grau de precisão. Analisados estatisticamente os resultados, ASHBURN positivou variações entre os professores dentro de cada grupo e, posteriormente, quando as dissertações voltaram a ser corrigidas pelos mesmos professores, os integrantes do grupo, além de variarem entre si, variaram, também, em relação aos resultados anteriores atribuídos por cada um deles. A pesquisa de ASHBURN concluiu, igualmente, que a aprovação ou reprovação de 40% dos sujeitos participantes da investigação dependeu não do que sabiam ou deixavam de saber, mas de *quem* leu os trabalhos. Finalmente, a pesquisa constatou que, em relação à aprovação ou reprovação de 10% dos sujeitos, o fator decisivo não foi o conhecimento ou o desconhecimento do assunto, mas sim o momento em que as dissertações foram lidas.

O problema da fidedignidade de notas atribuídas a dissertações elaboradas por uma amostra aleatória de 197 crianças, com a idade média de 12 anos, e possíveis candidatos à escola secundária, na Escócia, foi estudado empiricamente por FINLAYSON (1951). Os trabalhos foram corrigidos por 6 professores voluntários, com experiência de ensino na escola primária e na secundária. A correção baseou-se no método holístico (impressão geral). A pesquisa constatou que o desempenho dos professores, ao julgarem as dissertações, apresentou diferenças sensíveis quanto ao nível e à variabilidade das notas. A fidedignidade dos julgadores foi, em geral, baixa, apresentando uma correlação média, entre formas paralelas das dissertações, de 0,691. Ao recorrigirem as provas, dois meses depois, a consistência média foi alta (0,810), o que demonstrou a coerência dos julgadores no tempo. A fidedignidade das notas compostas, estimada pelo coeficiente de Spearman-Brown, também se apresentou elevada (0,880). A pesquisa de FINLAYSON estabeleceu correlações com vários critérios externos. Os coeficientes mostraram que a correlação mais alta (0,776) foi entre a dissertação e a estimativa do desempenho dos elementos do grupo por seus professores de Inglês, e as mais baixas com os resultados de um teste de QI (0,690); entretanto, as intercorrelações com a dissertação foram mais baixas do que as das outras variáveis entre si. FINLAYSON concluiu que isso decorreu da baixa fidedignidade das dissertações e porque, possivelmente, estas estariam medindo algo diferente do que mediam os outros instrumentos de medida, podendo esse "algo" ser, inclusive, erro de medida. A pesquisa, ao final, concluiu não haver suficiente prova para determinar se a dissertação deveria ou não substituir uma bateria de testes de seleção.

MYERS *et alii* (1966) concentraram sua pesquisa sobre fidedignidade na análise dos julgamentos de 145 professores, que, sob condições controladas, leram e corrigiram dissertações de 80.842 estudantes candidatos a Universidades. Os trabalhos, redigidos no prazo de 20 minutos, foram corrigidos, holística e independentemente, cada um deles por dois professores. A investigação de MYERS *et alii* procurou verificar, basicamente, se a fidedignidade dos julgadores se mantinha constante durante toda a correção.

Os coeficientes de fidedignidade variaram de 0,493, no terceiro dia, a 0,264, no quinto e último dia de correção. A fidedignidade neste último dia foi significantemente diferente dos coeficientes obtidos nos dias um, três e quatro. No segundo dia, a fidedignidade (0,364) não apresentou diferença significante em relação aos coeficientes dos demais dias. A pesquisa demonstrou não ter havido abaixamento da fidedignidade, exceto no último dia de correção. Tal fato teria resultado, possivelmente, de uma variação do nível de atenção dos julgadores, ao se aproximar o término dos trabalhos. Concluíram os pesquisadores que, independentemente da duração do período de correção, curto ou longo, sempre ocorre um abaixamento da fidedignidade no último momento. Finalmente, a pesquisa de MYERS *et alii* procurou verificar a influência do número de julgadores sobre a fidedignidade. Os coeficientes variaram de 0,406, para um julgador, a 0,732, para quatro avaliadores.

AKEJU (1972), na África, encontrou resultados semelhantes aos anteriormente apresentados, ao

analisar julgamentos de dissertações aplicadas pelo *West African Examinations Council.*

As fontes de erro que influenciam no julgamento de dissertações foram estudadas por FRENCH (1962), BRADDOCK *et alii* (1963) e, sobretudo, por COFFMAN (1971b, 1972). Este último, a partir de várias pesquisas experimentais, ressalta a inevitabilidade de três conclusões: 1.º — diferentes julgadores tendem a atribuir diferentes notas ao mesmo trabalho, 2.º — um único julgador tende a atribuir diferentes graus, em diferentes momentos, ao mesmo trabalho, e 3.º — as diferenças tendem a aumentar na medida em que as dissertações permitem grande liberdade de resposta. COFFMAN mostra, ainda, que essas conclusões revestem-se de maior complexidade do que aparentam. Assim sendo, analisa, inicialmente, o problema da variabilidade *entre julgadores* e conclui que os julgadores diferem quanto à severidade dos julgamentos, à maneira como distribuem os graus na escala adotada e, finalmente, diferem nos valores relativos que atribuem às dissertações. As implicações dessas fontes de erro, segundo COFFMAN (1971b), apesar de irrelevantes em situações de sala-de-aula, são, entretanto, de grande importância em programas externos de que participa um número elevado de indivíduos. Ao analisar a variabilidade *intra julgadores,* aponta três componentes de variação: o padrão relativo para diferentes trabalhos, o padrão geral de julgamento e a variabilidade dos julgamentos. A maior ou menor influência dessas fontes de erro depende do modo como os julgamentos serão utilizados (COFFMAN, 1971b). Se apenas para fins de classificação, a primeira fonte de erro deve constituir-se em motivo de preocupação; entretanto, se os resultados forem considerados medidas diretas de qualidade, então, todas as três fontes se tornam críticas.

O trabalho de COFFMAN (1971b) analisa problemas de erros de amostragem e ressalta o fato de que existe alguma evidência de incremento na fidedignidade por questão quando estas são "maiores", contudo, esse incremento não é proporcional ao tempo exigido para responder a questões maiores. Usando dados da pesquisa de GODSHALK *et alii* (1966), demonstra que a fidedignidade de um escore resultante de cinco julgamentos independentes de uma dissertação de 20 minutos seria de 0,485, enquanto que a fidedignidade do escore resultante da soma de dois escores, em duas dissertações, realizadas no mesmo período de tempo, seria de 0,655. Ao contrário, a fidedignidade de um escore baseado numa única dissertação de 40 minutos seria de somente 0,592. Assim, conclui COFFMAN que, em geral, quanto maior o número de diferentes dissertações incluídas no exame, maior a fidedignidade, pressupondo-se a mesma duração para a prova. Se o exame dissertativo tiver influência na determinação do status do estudante, COFFMAN destaca a necessidade da inclusão de uma amostra representativa de questões.

A partir de experiências quantitativas, COFFMAN (1971b) sugere alternativas para reduzir a influência dos erros de julgamento — uso de escalas de 7 a 15 unidades (mostra que, num conjunto de dissertações corrigidas numa escala de 15 pontos, a fidedignidade foi de 0,848; enquanto que, numa escala de 5 pontos, o mesmo coeficiente foi 0,767); definição clara e precisa das características dos vários pontos correspondentes na escala; correção de questão por questão e aplicação de várias provas aos mesmos sujeitos; e, finalmente, a realização de várias avaliações independentes para os mesmos elementos, porquanto a soma das avaliações é mais fidedigna do que uma única avaliação isolada.

Problemas relativos a aspectos diversos

O problema das flutuações da capacidade de expressão escrita do estudante foi pesquisado por BRADDOCK *et alii* (1963) e por McCOLLY (1970). Segundo BRADDOCK *et alii* nunca se pode estar certo de que o estudante esteja usando totalmente a sua capacidade e escrevendo tão bem quanto seria capaz. McCOLLY discute essa posição e, mesmo reconhecendo a influência dessas flutuações, mostra que, ao aplicar-se uma prova, cria-se, *ipso facto,* para o estudante, uma situação adversa, a que ele se deve ajustar para superar a problemática apresentada. Reconhece, entretanto, que essa posição é pessoal, com a qual muitos, possivelmente, não concordarão, tendo em vista a necessidade de avaliar-se um "melhor" desempenho do estudante. McCOLLY adverte que, a admitir-se tal fato, nunca serão alcançadas condições para avaliar o desempenho dos estudantes, pois a situação de exame é sempre artificial.

A influência da maior ou menor competência dos julgadores, em termos de nível de formação intelectual, sobre os graus atribuídos a dissertações, foi demonstrada experimentalmente por McCOLLY *et alii* (1963), com base na avaliação de 1.200 trabalhos dissertativos corrigidos por 16 professores de Inglês, com vários níveis de formação profissional. Por outro lado, DIEDERICH, *et alii* (1961), num experimento em que 300 dissertações foram corrigidas por professores de Inglês, Ciências Sociais, Ciências Naturais, advogados, administradores e editores, as fidedignidades variaram significantemente de acordo com os diferentes grupos ocupacionais.

As pesquisas sobre o número ideal de questões ou tópicos em provas de dissertação, além de raras, não oferecem dados suficientes e conclusivos (McCOLLY, 1970). Tendo em vista o fato de que o

desempenho do estudante varia de um assunto para outro, FRENCH (1962) condena o uso de um único tema em provas de dissertação; BRADDOCK *et alii* (1963) acham que, nesse tipo de prova, os estudantes devem escrever, no mínimo, sobre dois temas; entretanto, DIEDERICH *et alii* (1961) consideram que apenas dois assuntos são inteiramente inadequados para a medida da capacidade de expressão escrita. McCOLLY *et alii* (1963) usaram quatro temas em sua pesquisa, enquanto GODSHALK *et alii* (1966) estruturam o seu trabalho de pesquisa com base no desempenho de estudantes em cinco dissertações. O problema, ainda que não suficientemente pesquisado, é fundamental, sobretudo porque se relaciona com a validade de conteúdo da prova de dissertação.

A aparência externa das dissertações, ainda que possa parecer problema menor, é crítico no julgamento de dissertações, e constitui fonte tangível da falta de fidedignidade e de validade de dissertações (McCOLLY, 1970). O problema tem sido mais amplamente estudado que outros aspectos da dissertação; assim, McCOLLY cita inúmeros estudos de análise fatorial que indicam a emergência de um fator geral — apresentação — e de fatores específicos, como a caligrafia, que, segundo esses estudos, têm grande influência sobre o julgamento dos examinadores. CHASE (1968), sobre esse assunto, em estudo experimental, mostra que existe uma interação significante entre a qualidade da escrita e a ordem de leitura das dissertações, o que indica, possivelmente, o desenvolvimento progressivo de um viés negativo nos julgadores em relação à má caligrafia. Inversamente, muitos julgamentos superiores de dissertações parecem refletir um viés positivo e um prêmio para limpeza e legibilidade dos trabalhos.

A questão do método de correção adotado para o julgamento de provas de dissertação foi objeto de investigação por COFFMAN e KURFMAN (1968). A partir de dois experimentos, testaram a eficiência do método analítico e do holístico. Além desse objetivo, procuraram verificar, também, a possível influência da ordem seqüencial de apresentação dos trabalhos durante o julgamento. Estudaram, ainda, a influência da duração do período destinado à correção e, finalmente, investigaram a questão da utilização de diferentes padrões de correção.

A análise de variância dos resultados mostrou não haver diferenças significantes atribuíveis aos métodos de correção. O problema da igual eficiência dos métodos holísticos e analítico está sujeito, entretanto, a divergências. A despeito dos resultados de COFFMAN e KURFMAN, e de COFFMAN (1971a) aconselhar o emprego do método geral (holístico) que, na sua opinião, proporcionará resultados mais fidedignos, no estudo de DIEDERICH *et alii* (1961) foi observado que os professores tendem a considerar diferentes qualidades no trabalho, isto é, realizam uma correção analítica. A análise fatorial procedida por DIEDERICH *et alii* indicou cinco fatores: idéias, forma, estilo, mecânica e fraseado *(wording)*. PAGE (1968), em experimento com 256 dissertações, usou as cinco categorias identificadas por DIERERICH *et alii* e concluiu que, no julgamento de múltiplos traços, existe o perigo da influência do efeito de halo. Ou seja, que o julgador, ao se manifestar sobre uma determinada qualidade, esteja, na verdade, respondendo a uma impressão geral decorrente da qualidade da dissertação. PAGE, entretanto, não rejeita a correção pelo método analítico e, com base na análise da variância, mostra a ocorrência de uma interação traço-dissertação significante, que traduziria alguma "validade" dos diferentes julgamentos dos traços.

A pesquisa de COFFMAN e KURFMAN (1968) mostrou não existir diferenças significantes quando as dissertações, para fins de correção, são distribuídas entre os julgadores de modo aleatório, o mesmo não ocorrendo quando, durante as várias correções, a ordem de distribuição é fixa. Ao considerarem as variáveis tempo e padrão de julgamento, a pesquisa constatou diferenças significantes. Os julgadores, nos experimentos em questão, tenderam a atribuir escores mais baixos no segundo dia de trabalho do que no primeiro, e a fidedignidade dos resultados mostrou-se comprometida, em virtude da variação dos padrões de correção.

As pesquisas sobre o número de pontos ou intervalos em escalas para julgamento de dissertações ainda são em número restrito. McCOLLY (1965) estudou a possível diferença nas distribuições de julgamentos quando são utilizadas escalas de 4 e de 6 pontos. Ainda que não tenham sido positivadas diferenças, McCOLLY constatou que os julgadores eram significantemente mais morosos nesta última escala do que na de quatro pontos. COFFMAN (1971a), ao contrário, verificou que, em certas situações, os julgadores são tão rápidos em escalas longas quanto em escalas curtas. McCOLLY (1970) chama a atenção para o fato de que numerosos problemas relativos à natureza das escalas para correção de dissertações ainda permanecem no campo da especulação teórica.

MARSHALL e POWERS (1969) pesquisaram a influência de fatores estranhos na correção de dissertações, com base no julgamento de 420 futuros professores sobre o conteúdo de um trabalho dissertativo. As instruções para os julgadores destacaram o fato de que apenas o aspecto conteudístico deveria ser levado em consideração no julgamento. A análise dos dados levantados por MARSHALL e POWERS mostrou que os avaliadores, apesar da advertência limitativa, foram influenciados pela

qualidade da composição (erros de ortografia, gramática e pontuação) e pela legibilidade da escrita.

Problemas relativos à validade

A pesquisa de GODSHALK *et alii* (1966), sem dúvida o estudo mais completo, até a presente data, sobre a medida da capacidade de expressão escrita, apresenta um conjunto de análises comparativas referentes ao valor de itens objetivos, exercícios interlineares e de redações na medida da expressão escrita. Inicialmente, a pesquisa procurou estabelecer a validade de dissertações realizadas num período de 20 minutos, assim como a de outros tipos de questões; aos poucos, entretanto, a pesquisa que reúne estudos realizados entre 1945 e 1960, modificou os seus objetivos iniciais e realizou uma análise compreensiva sobre os problemas relacionados com a medida da capacidade de expressão escrita.

GODSHALK *et alii* utilizaram oito testes experimentais e cinco dissertações, que foram aplicados a uma amostra de estudantes da escola média *(high school juniors e seniors)*, e estabeleceram as relações entre os escores nos testes e as classificações nas dissertações. Foram utilizadas seis classes de itens objetivos, para diferentes fins: 1 — *itens de uso* (reconhecimento de uso defeituoso, inclusive gramática, dicção, estrutura básica e mecânica), 2 — *correção de sentença* (seleção da melhor forma de uma determinada parte da frase), 3 — *organização de parágrafo* (estruturação de várias sentenças num parágrafo coerente), 4 — *grupo de prosa* (itens baseados num parágrafo com uma sentença omitida e que deve ser identificada), 5 — *identificação de erro* (reconhecimento da existência ou não de determinados erros) e 6 — *mudança de estrutura* (decisão pelo examinando de alterações na sentença quando um elemento é modificado de determinada forma). (Os itens 1, 2, 5 e 6 foram adaptados pela Fundação Carlos Chagas, em 1970, e são utilizados em provas de Comunicação e Expressão, em Concursos Vestibulares).

Além dos itens anteriormente descritos, a pesquisa empregou exercícios interlineares, os quais consistem em textos mal elaborados, cujas deficiências devem ser identificadas e corrigidas pelos examinandos. A pesquisa utilizou dois tipos de textos: um, narrativo; outro, expositivo. Tal procedimento visou a compensar possíveis deficiências de compreensão do material pelos examinandos.

A fim de coletar elementos para o critério de validação, aplicaram cinco dissertações, das quais duas exigiam tratamento mais elaborado, inclusive análise e interpretação, sendo as mesmas redigidas em dois períodos de 40 minutos. As outras três dissertações constaram da redação de simples parágrafos, escritos em sessões de 20 minutos. A temática das dissertações exigiu dos examinandos diferentes comportamentos: 1 — *descrição* (Elemento interessante da cidade natal descrito a um amigo estrangeiro), 2 — *narrativa* (Estória baseada em experiência pessoal), 3 — *exposição* (A favor ou contra a idéia de que os adolescentes seriam mais conservadores do que os adultos), e 4 — *argumentação* (dois ensaios: um, sobre a atitude a tomar relativamente ao comportamento negligente de um estudante; outro, análise das idéias de um universitário calouro durante uma palestra para estudantes do seu antigo ginásio).

O material descrito foi aplicado a duas turmas dos dois últimos anos de 24 escolas secundárias, públicas e particulares, que, em média, apresentavam 55 estudantes para cada uma delas. Obtiveram-se dados completos para 646 casos. Além dos dados coletados pelos instrumentos descritos, os pesquisadores analisaram o desempenho de 533 sujeitos no PSAT *(Preliminary Scholastic Aptitude Test)* e de 158 no SAT *(Scholastic Aptitude Test)*.

Os escores dos testes e dos exercícios interlineares foram obtidos segundo procedimento padronizado. Os exercícios interlineares foram avaliados por 25 professores treinados para esse fim e supervisionados por um especialista. Os escores destes exercícios e os dos seis testes objetivos constituíram um conjunto de oito variáveis que foram correlacionadas separadamente e combinadas de vários modos com os escores do critério. Os mesmos julgadores dos exercícios interlineares corrigiram, também, as dissertações, o que foi feito no período de cinco dias. A correção utilizou o método holístico ou global e cada dissertação foi classificada em três níveis: 3 (superior), 2 (médio) e 1 (deficiente), sem maiores preocupações com uma possível distribuição normal dos resultados. Os trabalhos foram corrigidos independentemente por cinco julgadores; por outro lado, os pesquisadores promoveram uma distribuição e redistribuição das dissertações de forma que todos os 25 professores tivessem a oportunidade de corrigir pelo menos uma dissertação de cada aluno.

Após a análise dos resultados, a equipe de GODSHALK chegou às seguintes amplas generalizações:

1.º — *A fidedignidade dos escores de dissertação depende sobretudo do número de diferentes dissertações e do número de diferentes julgamentos.* No caso, com cinco dissertações e cinco diferentes julgadores para cada trabalho, a fidedignidade dos julgadores foi 0,92 e a dos escores aproximadamente 0,84. Ao contrário, para uma única dissertação e um único julgador, os mesmos coeficientes foram, respectivamente, 0,40 e 0,25. A fidedignidade de quatro julgadores para uma única dissertação apresentou um coeficiente de 0,70.

2.º — *Quando questões objetivas, especialmente planejadas para medir a capacidade de expressão escrita, são avaliadas em relação a um critério fidedigno da mesma capacidade, demonstram ser altamente válidas.* Para um total de 646 casos, a pesquisa estimou que se cinco dissertações adicionais, sobre tópicos paralelos, fossem julgadas do mesmo modo, o escore total apresentaria uma correlação de 0,84 com o critério. Os coeficientes de validade, para os diferentes tipos de questões objetivas, variaram de 0,568 a 0,709. Os escores compostos dos três diferentes subtestes apresentaram correlação superior a 0,70 e a mais elevada chegou a 0,755. Quando os escores do PSAT foram incluídos, as correlações múltiplas foram ainda maiores, mas o aumento foi pequeno.

3.º — *O preditor mais eficiente de uma medida fidedigna da capacidade de expressão escrita inclui questões de dissertação ou exercícios interlineares combinados com questões objetivas.* Os coeficientes de validade para os exercícios interlineares não foram tão altos quanto os dos subtestes objetivos, mas se situaram nos limites da amplitude geral; por outro lado, quando combinados com os subtestes objetivos, os exercícios interlineares, em geral, apresentaram os mais altos coeficientes de validade. Quanto às dissertações, as validades somente se aproximaram da amplitude de validade para outros tipos de questões quando os escores se basearam em três julgamentos. A combinação dos escores da dissertação com os escores dos subtestes objetivos determinou coeficientes de validade maiores do que os obtidos por meio de outras combinações, que incluíam exercícios interlineares.

O problema da validação de testes objetivos, usando a dissertação como critério, foi estudado por COFFMAN (1966), que, a partir dos dados da pesquisa de GODSHALK *et alii* (1966), submeteu os elementos por estes levantados a novos tratamentos estatísticos e demonstrou, por meio do controle, ora do número de dissertações ora do número de julgadores, que a fidedignidade das dissertações pode ser aumentada desde que se aumente o número de questões ou o número de julgadores independentes. A partir de pressupostos psicométricos, COFFMAN construiu uma tabela de diferentes coeficientes de validade que podem ser esperados quando o número de temas e o número de avaliadores das dissertações variam. O trabalho de COFFMAN destaca o fato de que, para obter coeficientes de validade altos, há necessidade de o estudante escrever, no mínimo, sobre dois temas, que devem ser corrigidos independentemente por cinco julgadores. Ou, então, que disserte sobre três tópicos, que devem ser submetidos à correção independente de três julgadores.

O trabalho apresentado por PALMER (1961), sobre o problema da dissertação, ainda que não possa ser considerado estritamente experimental, baseia-se, contudo, em estatísticas levantadas pelo *College Entrance Examination Board (English Committees)* em pesquisas com testes objetivos e dissertações. Após examinar as seguintes variáveis: tarefas solicitadas, raciocínios e compreensão exigidos, flutuações da dificuldades, fidedignidade dos escores, dos julgadores e dos examinandos (variações na capacidade de expressão escrita em função do tema), e validade das provas, em função dos critérios — graus em cursos, classificações feitas por professores e fator verbal no teste SAT —, PALMER concluiu que as dissertações não são fidedignas e nem válidas e que, apesar de suas falhas, os testes (refere-se aos testes de Inglês elaborados para o CEEB) constituem um método válido e fidedigno de investigar a capacidade de expressão escrita do estudante.

REFERÊNCIAS BIBLIOGRÁFICAS

AKEJU, A. A. (1972) — The reliability of General Certificate of Education Examination. English Composition Papers in West Africa. *Journal of Educational Measurement*, 9,3.

ASHBURN, R. R. (1938) — An experiment in the essay-tipe question. *Journal of Experimental Education*, 7.

BARROSO, C. L. de M. (1972) — Estudos de predição do comportamento acadêmico: II — Faculdades de Medicina. Fundação Carlos Chagas. *Cadernos de Pesquisa*, nº 5.

BARROSO, C. L. de M. (1972) — Pesos nominais e pesos efetivos no Vestibular do CESCEM. Fundação Carlos Chagas. *Cadernos de Pesquisa*, nº 6.

BARROSO, C. L. de M. *et alii* (1972) — Estudos de predição do comportamento acadêmico: I — Faculdade de Medicina Veterinária. Fundação Carlos Chagas. *Cadernos de Pesquisa*, nº 5.

BESSA, N. M. (1976) — *University entrance examinations in Brazil: measurement procedures.* (Paper presented at the Symposium of Measurement Procedures by Various Countries in the Selection of College Students. AERA. Washington, 1975) Fundação Carlos Chagas.

BRADDOCK, R. *et alii* (1963) — Research in written compositions, *National Council of Teachers of English.* Champaign, Ill, in McColly, 1970.

BREEN, T. F. (1975) — Estabilidade do concurso vestibular do CESCEM. Fundação Carlos Chagas. *Cadernos de Pesquisa*, nº 12.

CHASE, C. I. (1968) — The impact of some obvious variables on essay test scores. *Journal of Educational Measurement*, 3.2.

COFFMAN, W. E. (1966) — On the validity of essay tests of achievement. *Journal of Educational Measurement*, 3.2.

COFFMAN, W. E. (1971a) — Essay examination, in Thorndike, R. L. (ed) *Educational Measurement.* Washington, D. C. American Council on Education.

COFFMAN, W. E. (1971b) — On the reliability of ratings of essay examinations in English. *Research in the Teaching of English,* 5.1.

COFFMAN, W. E. (1972) — On the reliability of ratings of essay examination. *Measurement in Education,* 3.3.

COFFMAN, W. E. e KURFMAN, D. (168) — A comparison of two methods of reading essay examination. *American Educational Research Journal,* 5.1.

DIEDERICH, P. *et alii* (1961) — *Factors in Judgement of Writing Ability.* Princeton, New Jersey — Educational Testing Service, in McColly, 1970.

FINLAYSON, D. S. (1951) — The reliability of marking essays. *British Journal of Educational Psychology,* 21.

FRENCH, J. W. (1962) — Schools of thought in judging excellence of English Themes. *Proceedings of Invitational Conference on Testing Problems.* Princeton, New Jersey. Educational Testing Service, in McColly, 1970.

GODSHALK, F. I. *et alii* (1966) — *The measurement of writing ability.* New York. College Entrance Examination Board.

McCOLLY *et alii* (1963) — Comparative Effectiveness of Composition Skills Learning Activities. U. S. Office of Education. *Cooperative Research Project 1528.* The University of Wisconsin, Madison, in McColly, 1970.

McCOLLY *et alii* (1965) — Composition Ratings Scales for General Merit: an experimental evaluation. *Journal of Educational Research,* 59, in McColly, 1970.

McCOLLY, W. (1970) — What does research say about the judging of writing ability. *Journal of Educational Research,* 64.

MARSHALL, J. C. e POWER, J. M. (1969) — Writing neatness, composition errors and essay grades. *Journal of Educational Measurement,* 6.

MORAES, R. e ANDRADE, E. M. (1970) — Análise das provas do vestibular. PUC. São Paulo.

MORAES, R. e ANDRADE, E. M. (1971) — Caracteristicas psicológicas de universitários. *Revista da Pontifícia Universidade Católica de São Paulo,* Vol. XLIV, 87/88.

MYERS, A. E. *et alii* (1966) — Simplex structure in the grading of essay tests. *Educational and Psychological Measurement,* 26.

PAGE, E. B. (1968) — The analysis of essay by computer. U. S. Office of Education. *Cooperative Research Project* 1318. The University of Connecticut, Storrs.

PALMER, D. (1961) — Sense or Nonsense. The objective testing of English Composition. *The English Journal,* 50.

RODRIGUES, A. (1972) — *Análise técnica das provas do vestibular,* 72. Fundação Cesgranrio. Rio de Janeiro.

RODRIGUES, A. (1974) — *Testes de aptidão na seleção de candidatos ao ensino superior.* Relatório Técnico nº 1. Fundação Cesgranrio. Rio de Janeiro.

SANCHEZ, V. F. (1972) — Um estudo de fidedignidade da Taxonomia dos objetivos educacionais: — domínio cognitivo. Fundação Carlos Chagas. *Cadernos de Pesquisas,* nº 6.

STARCH, D. e ELLIOTTE, E. C. (1972) — Reliability of grading of high-school work in English. *School Review,* 20.

STARCH, D. e ELLIOTT, E. C. (1913a) — Reliability of grading work in History. *School Review,* 21.

STARCH, D. e ELLIOTT, E. C. (1913b) — Reliability of grading work in Mathematics. *School Review,* 21.

[Recebido para publicação em dezembro de 1975]