CONFIABILIDADE E CONCORDÂNCIA ENTRE JUÍZES: APLICAÇÕES NA ÁREA EDUCACIONAL

DANIEL ABUD SEABRA MATOS

RESUMO

Os objetivos desta pesquisa foram: (1) investigar as estratégias de verificação da confiabilidade e concordância entre juízes, enfatizando as aplicações na área educacional; (2) realizar uma revisão da literatura nacional sobre as técnicas de confiabilidade e concordância entre juízes e suas áreas de aplicação; e (3) ilustrar a aplicação das técnicas de confiabilidade e concordância entre juízes por meio da análise das correções das redações do vestibular de uma universidade pública de Minas Gerais. Utilizamos o coeficiente de correlação intraclasse para analisar a confiabilidade e concordância entre juízes na correção das redações no período de 2005 a 2010. Identificamos pouco uso, nas pesquisas educacionais, de técnicas de concordância entre juízes. Quanto à análise da correção das redações, alguns resultados foram satisfatórios (exemplo: confiabilidade média dos juízes para as notas totais das redações) e outros insatisfatórios (exemplo: concordância baixa em alguns critérios de correção).

PALAVRAS-CHAVE TAXA DE CONFIABILIDADE • CONCURSO VESTIBULAR • JUÍZES • REDAÇÃO.

RESUMEN

Los objetivos de la presente investigación fueron los siguientes: (1) investigar las estrategias de verificación de la confiabilidad y concordancia entre jueces, enfatizando las aplicaciones en el área educativa; (2) realizar una revisión de la literatura nacional sobre las técnicas de confiabilidad y concordancia entre jueces y sus áreas de aplicación; y (3) ilustrar la aplicación de las técnicas de confiabilidad y concordancia entre jueces por medio del análisis de las correcciones de las redacciones del examen de ingreso a una universidad pública de Minas Gerais. Utilizamos el coeficiente de correlación intraclase para analizar la confiabilidad y concordancia entre jueces en la corrección de las redacciones en el periodo de 2005 a 2010. Identificamos poco uso, en las investigaciones educativas, de técnicas de concordancia entre jueces. En lo que se refiere al análisis de la corrección de las redacciones, algunos de los resultados fueron satisfactorios (ejemplo: confiabilidad media de los jueces para las notas totales de las redacciones) y otros insatisfactorios (ejemplo: baja concordancia en algunos criterios de corrección).

PALABRAS CLAVE TASA DE CONFIABILIDAD • EXAMEN PARA INGRESO A LA UNIVERSIDAD• JUECES • REDACCIÓN.

ABSTRACT

The aims of this study were to: (1) investigate the strategies for verifying reliability and agreement among evaluators, focusing on the applications in the educational area; (2) conduct a review of the national literature on the techniques of reliability and agreement among judges and their areas of application; and (3) illustrate the application of the techniques of reliability and agreement among evaluators by analyzing the corrections of the Vestibular (college entrance exam) essays from one public university in Minas Gerais. We used the intraclass correlation coefficient to analyze the reliability and agreement among evaluators in the correction of the essays from 2005 to 2010. We identified little use, in the educational research, of agreement techniques among evaluators. As for the analysis of the correction of essays, some results were satisfactory (example: mean reliability of the evaluators for total scores of the essays) and others were unsatisfactory (example: low agreement in some criteria of correction).

KEYWORDS RATE OF RELIABILITY • VESTIBULAR EXAMINATION • EVALUATORS • ESSAY.

INTRODUÇÃO

Nos últimos anos, tem acontecido um aumento do interesse na área de avaliação e suas aplicações no campo educacional. Esse contexto produziu um crescimento significativo de pesquisas sobre a avaliação educacional. Também tem se configurado atualmente um maior compromisso do governo, das universidades, das escolas e dos profissionais da educação com a realização de avaliações confiáveis e de boa qualidade.

Algumas possíveis definições de avaliação incluem tanto definições mais gerais quanto definições mais específicas do campo educacional. São alguns exemplos: um processo de delineamento, obtenção e fornecimento de informações que permitam julgar alternativas de decisão (STUFFLEBEAM, 1971) e um procedimento sistemático e compreensivo em que se utilizam estratégias diversas para avaliar a trajetória acadêmica e pessoal do estudante (QUINTANA, 2003). Ainda com relação à definição de avaliação, para Jorba e Sanmartí (2003), toda atividade de avaliação é um processo com três etapas: 1) coleta de informação, que pode ser ou não instrumentada; 2) análise dessa informação e conclusão sobre o

resultado da análise; e 3) tomada de decisões de acordo com a conclusão.

Além disso, para cumprir seus propósitos, a avaliação precisa atender a alguns requisitos. Assim, para que uma avaliação seja considerada de boa qualidade precisa ter, por exemplo, validade (*validity*) e confiabilidade (*reliability*). Validade pode ser definida como "o grau em que todas as evidências acumuladas corroboram a interpretação pretendida dos escores de um teste para os fins propostos" (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 1999). Já a confiabilidade está relacionada com a consistência e precisão dos resultados do processo de mensuração (URBINA, 2007).

Portanto, dentre as várias possibilidades de abordagem da avaliação (inclusive as diferentes etapas do processo avaliativo), podemos afirmar que o principal foco desta pesquisa envolve a etapa de análise da informação e dos resultados, com ênfase no critério da confiabilidade. Mais especificamente, investigamos estratégias de verificação da confiabilidade e concordância entre juízes (*inter-rater agreement and reliability*) e suas aplicações na área educacional. O termo juiz é usado aqui como sinônimo de avaliador, ou seja, em situações nas quais mais de um juiz (avaliador) participa de um processo avaliativo, existem estratégias para verificar qual o grau de concordância desses juízes (avaliadores). Em situações como essas, é crucial verificar se existe um nível mínimo de concordância entre os juízes. A importância disso reside no fato de explicitar a confiabilidade de um processo avaliativo. Se dois ou mais juízes discordam muito em uma avaliação, isso pode indicar uma falta de confiabilidade nos resultados.

A literatura especializada aponta diversas maneiras de medir o nível de confiabilidade e concordância entre juízes, incluindo técnicas como porcentagem, correlação, coeficiente *Kappa* de Cohen, dentre outras (HANEY et al., 2004). No entanto, existe uma lacuna nas pesquisas educacionais brasileiras, pois estratégias de verificação do nível de confiabilidade e concordância entre juízes têm sido pouco estudadas e aplicadas no campo educacional. Essa, porém, parece ser uma realidade diferente da área de ciências da saúde no

Brasil, em que se verifica uma maior aplicação das técnicas de confiabilidade e concordância entre juízes (ANDRADE; SHIRAKAWA, 2006; BRUSCATO; IACOPONI, 2000; DEL-BEN et al., 2001; FRAGA-MAIA; SANTANA, 2005; PERROCA; GAIDZINSKI, 2002, 2003; POLANCZYK et al., 2003; VENTURA; BOTTINO, 2001).

Em face da relevância dessa área de trabalho, realizamos a presente pesquisa cujos objetivos foram: (1) investigar as estratégias de verificação da confiabilidade e concordância entre juízes, enfatizando as aplicações na área educacional; (2) realizar uma revisão da literatura nacional sobre as técnicas de confiabilidade e concordância entre juízes e suas áreas de aplicação (ciências humanas e ciências da saúde); e (3) ilustrar a aplicação das técnicas de confiabilidade e concordância entre juízes por meio da análise das correções das redações do vestibular de uma universidade pública do estado de Minas Gerais.

CONFIABILIDADE E CONCORDÂNCIA ENTRE JUÍZES

Pesquisadores e profissionais geralmente usam o termo confiabilidade entre juízes (*inter-rater reliability*) como uma expressão genérica para a consistência entre avaliadores. No entanto, alguns especialistas em avaliação adotam uma definição mais precisa para o termo. A confiabilidade entre juízes pode ser, assim, definida como uma medida da consistência entre avaliadores na ordenação ou posição relativa de avaliações de desempenho, independentemente do valor absoluto da classificação de cada avaliador. Já a concordância entre juízes (*inter-rater agreement*) pode ser definida como o grau em que dois ou mais avaliadores, utilizando a mesma escala de avaliação, fornecem igual classificação para uma mesma situação observável. Dessa maneira, ao contrário da confiabilidade entre juízes, a concordância entre juízes é uma medida da consistência entre o valor absoluto das classificações dos avaliadores (GRAHAM et al, 2012). Além disso, é possível que dois avaliadores tenham pouca ou nenhuma concordância e ainda assim apresentem uma confiabilidade alta (TINSLEY; WEISS, 2000). O Quadro 1 ilustra essa possibilidade:

QUADRO 1 – Diferença entre confiabilidade e concordância

	CONCORDÂNCIA BAIXA, CONFIABILIDADE ALTA		CONCORDÂNCIA ALTA, CONFIABILIDADE ALTA	
	JUIZ 1	JUIZ 2	JUIZ 3	JUIZ 4
PROFESSOR A	1	2	1	1
PROFESSOR B	2	3	2	2
PROFESSOR C	3	4	3	3
PROFESSOR D	4	5	4	4
CONCORDÂNCIA	0,0		1,0	
CONFIABILIDADE	1,0		1,0	

Fonte: Tinsley, Weiss (2000).

A concordância mede com que frequência dois ou mais avaliadores atribuem exatamente a mesma classificação. A confiabilidade mede a semelhança relativa entre dois ou mais conjuntos de classificações. Nesse sentido, o Quadro 1 exemplifica a diferença entre confiabilidade e concordância. Os juízes 1 e 2 estão de acordo sobre o desempenho relativo dos quatro professores, pois ambos atribuíram classificações que aumentam gradativamente (o professor A recebe o menor escore e o Professor D recebe o maior escore). No entanto, embora eles concordem sobre o ranqueamento relativo dos quatro professores, não concordaram nenhuma vez sobre o nível absoluto de desempenho. Consequentemente, o nível de confiabilidade entre os juízes 1 e 2 foi perfeito (1.0), mas não existiu nenhuma concordância. Já os juízes 3 e 4 concordaram tanto sobre o nível absoluto quanto sobre a ordem relativa do desempenho dos professores. Dessa forma, eles tiveram a confiabilidade (1.0) e a concordância perfeitas (1.0) entre juízes (GRAHAM et al., 2012).

Uma das possíveis explicações para a diferença entre confiabilidade e concordância é a utilização de pontos de ancoragem distintos. Por exemplo: juízes realizando classificações em uma escala de 1 a 10. Imaginemos que o juiz 1 avalia todos os sujeitos com escores altos na parte superior da escala (entre 5 e 10) e o juiz 2 avalia todos com escores baixos (entre 1 e 5 na escala). Numa situação como essa, precisamos definir se a variabilidade individual do juiz é

importante para os fins de uma dada pesquisa. Se não for, simplesmente testamos se cada juiz classificou cada observação de uma maneira semelhante (ordenação ou posição relativa de avaliações – consistência/confiabilidade). Mas se a variabilidade individual do juiz for importante, então testamos se cada juiz deu para cada observação exatamente o mesmo escore (diferença no valor absoluto – concordância). Assim, nessa situação, podemos encontrar alta consistência/confiabilidade e pouca ou nenhuma concordância entre os juízes.

Normalmente, a concordância entre juízes é mais importante para os educadores quando eles tomam decisões de alto impacto (*high-stakes decisions*), como retenção ou promoção. Isso ocorre porque, muitas vezes, precisamos tomar decisões com base num limiar de pontuação com um critério de corte. A concordância entre juízes também é importante quando informa os resultados de avaliação com o intuito de fornecer *feedback*. A confiabilidade entre juízes é mais frequentemente utilizada em pesquisas ou onde o único interesse é na consistência das decisões dos avaliadores sobre os níveis relativos de desempenho. Com base nessas definições, a concordância entre juízes pode ser considerada a medida de maior interesse para avaliações educacionais (GRAHAM et al., 2012).

Ainda com relação à nomenclatura empregada na literatura, destacamos que neste trabalho usamos preferencialmente a expressão "concordância entre juízes". No entanto, quando necessário, efetuamos a distinção entre os conceitos.

Quanto aos métodos para calcular a concordância entre juízes, a literatura indica vários, sendo que a porcentagem de concordância absoluta (*percentage of absolute agreement*) é a técnica mais simples utilizada. Ela consiste unicamente em calcular o número de vezes em que os avaliadores concordam e dividir pelo número total de avaliações (varia entre 0 e 100%). Para Stemler (2004), o valor de 75% é considerado o mínimo de concordância aceitável, já valores a partir de 90% são considerados altos. Uma desvantagem dessa técnica reside no fato de ela não levar em consideração a proporção de concordância devido ao acaso.

Nesse sentido, as tentativas de estimativas melhores da concordância entre juízes começaram com o desenvolvimento

do coeficiente *kappa* de Cohen – K (*Cohen's kappa coefficient*) (COHEN, 1960). O coeficiente *kappa* é um procedimento estatístico que leva em consideração no seu cálculo a probabilidade de concordância ao acaso (CROCKER; ALGINA, 2009). Assim, esse coeficiente pode ser definido como a proporção de concordância entre os juízes após ser retirada a proporção de concordância devido ao acaso (FONSECA et al., 2007). O *kappa* varia entre 0 e 1, podendo ser interpretado da seguinte forma: K<0,4 é pobre; 0,4≤K<0,75 é satisfatório a bom; K≥0,75 é excelente (FLEISS, 1981). No entanto, esse critério de corte não é totalmente consensual na literatura. Um dos primeiros critérios foi o proposto por Landis e Koch (1977): K< 0: sem concordância; 0≤K<0,21: presença de ligeira concordância; 0,21≤K<0,41: concordância fraca; 0,41≤K<0,61: concordância moderada; 0,61≤K<0,81: concordância substancial; 0,81≤K≤1,00: concordância quase perfeita. Altman (1991) considera a concordância alta a partir de 0,80. Destacamos, ainda, que a literatura descreve o *kappa* como o método mais utilizado quando as variáveis são nominais (FONSECA et al., 2007).

Para variáveis ordinais, existe o chamado *kappa* ponderado (kw), que permite atribuir pesos diferentes às concordâncias e aos desacordos. Assim, esse coeficiente, desenvolvido por Cohen (1968), é um índice de concordância preferível quando classificamos um conjunto de dados em categorias ordenadas, pois o k distingue apenas entre acordo e desacordo em categorias nominais (SCHUSTER, 2004). Em 1981, Fleiss elaborou uma extensão do *kappa* e a denominou *Fleiss' generalized kappa*, para incluir casos em que existem três ou mais juízes (KING, 2004). Dessa maneira, a limitação do *kappa* original (medir a concordância apenas entre dois juízes) foi superada pela inclusão de múltiplos juízes.

Nos últimos anos, existe uma preocupação crescente com relação às limitações do *kappa*. As principais críticas são: o k é afetado pela distribuição heterogênea de categorias (problema de prevalência) e pela extensão na qual os juízes discordam (problema de viés) (BLOOD; SPRATT, 2007). Como uma forma de superar essas limitações, Gwet (2001) desenvolveu duas novas estatísticas: *first-order agreement coefficient* (AC1)

e *second-order agreement coefficient* (AC2) (GWET, 2001). O AC1 é utilizado com dois ou mais juízes e uma escala de classificação com duas ou mais categorias. Já o AC2 também é utilizado com dois ou mais juízes, mas com uma escala de classificação ordenada contendo duas ou mais categorias (BLOOD; SPRATT, 2007). Assim como o *kappa*, ambos os coeficientes AC1 e AC2 variam entre 0 e 1 e possuem a mesma interpretação: quanto mais próximo de 1, melhor (menor a probabilidade de a concordância acontecer devido ao acaso). Entretanto, as pesquisas com as estatísticas AC1 e AC2 ainda são poucas e as conclusões baseadas nesses índices ainda precisam ser cautelosas (BLOOD; SPRATT, 2007). Dessa forma, o coeficiente *kappa* continua sendo largamente utilizado para aferir a concordância entre juízes.

Para dados contínuos, existe um tipo especial de coeficiente de correlação denominado intraclasse – CCI (*intra-class correlation coefficient*), que é a medida de concordância mais utilizada para variáveis contínuas (LU, 2007). A correlação de Pearson mede a intensidade da associação interclasse (entre variáveis de classes diferentes, entre construtos diferentes). Já o coeficiente de correlação intraclasse mede essa intensidade dentro de uma mesma classe (diferentes medidas de um mesmo construto), que podem ser medidas repetidas de um mesmo participante ou medidas de várias pessoas dentro de um mesmo grupo (exemplo: estudantes em uma sala de aula). Assim, o CCI é aplicado em dados estruturados em grupos, sendo obtido dividindo o valor da variação entre os indivíduos pela variação total. O coeficiente de correlação intraclasse é uma medida de concordância corrigida pela concordância esperada ao acaso (BLAND; ALTMAN, 1990). Além disso, existem tipos diferentes de CCI e ele pode ser estimado de diversas formas. Alguns tipos de CCI são bastante conhecidos na literatura, o Coeficiente Alfa de Cronbach, por exemplo, é muito utilizado na área de psicometria para avaliar a consistência interna (confiabilidade), a similaridade entre os itens de um teste.

Uma das vantagens do CCI é que ele representa a concordância entre dois ou mais juízes ou entre várias medidas feitas pelo mesmo juiz. O CCI é equivalente à estatística

kappa para variáveis contínuas: também varia entre 0 e 1, podendo ser interpretado da seguinte forma: CCI<0,4 é pobre; 0,4≤CCI<0,75 é satisfatório a bom; CCI ≥0,75 é excelente (FLEISS, 1981). Esses critérios também podem variar na literatura, inclusive dependendo da técnica empregada. Para Hair *et al*. (2005), por exemplo, os valores de Alfa de Cronbach de 0,60 a 0,70 são considerados como limite inferior da aceitabilidade. Entretanto, segundo Graham *et al*. (2012), existe pouco consenso sobre um valor suficiente para o CCI. Enquanto 0,70 pode ser suficiente para uma medida utilizada para fins de pesquisa, alguns pesquisadores defendem um valor de 0,8 a 0,9 como o mínimo para a tomada de decisões importantes sobre pessoas (HAYS; REVICKI, 2005).

Destacamos, ainda, que a literatura aponta outros métodos para verificar o nível de concordância entre juízes. São exemplos disso: o coeficiente *Bennett's S,* a Teoria de Resposta ao Item (TRI) (EMBRETSON; REISE, 2000), dentre outros. Primi *et al*. (2007), por exemplo, analisaram o nível de confiabilidade de juízes com relação ao Teste de Criação de Metáforas (TCM). Esse teste é composto por nove itens, aos quais os participantes deram um total de 513 respostas. Cada resposta foi avaliada de forma independente pelos juízes, em uma escala de 0 a 3, correspondente ao nível de elaboração da metáfora. A concordância foi aferida por meio do modelo de Rasch, assumindo cada ideia como sendo um caso e cada juiz como sendo um item de um teste hipotético. Os autores utilizaram um procedimento indicado por Linacre (1998), denominado rede de juízes ancorados (*judge-linking network*). Na matriz de dados, foi aplicado o modelo de Rasch de créditos parciais (PRIMI et al., 2007).

MÉTODO

Inicialmente, realizamos uma busca na literatura brasileira sobre trabalhos que abordam as técnicas de verificação do nível de concordância entre juízes. Bancos de dados do *Scientific Electronic Library Online* (SciELO) e da Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) foram pesquisados na internet. Combinamos os seguintes descritores usando

tanto o singular quanto o plural: nível, concordância, confiabilidade, juízes, avaliadores. Também utilizamos descritores em inglês (*inter-rater, intercoder, agreement, realiability*), pois alguns periódicos nacionais publicam artigos em inglês. Em alguns casos, foram encontrados somente os resumos dos estudos, que não foram incluídos nessa revisão da literatura porque não atendiam ao critério de fornecer dados suficientes para a análise das pesquisas. Nossa revisão sobre as técnicas de concordância entre juízes é analisada pela área de aplicação (ciências da saúde ou ciências humanas) e pela identificação dos métodos mais utilizados.

Após o levantamento bibliográfico, ilustramos a aplicação das técnicas de confiabilidade e concordância entre juízes por meio da análise das correções das redações do vestibular de uma universidade pública do estado de Minas Gerais. Cada redação foi corrigida por dois avaliadores. No caso de discrepância nas notas, um terceiro avaliador participava da correção. Analisamos todas as redações da segunda etapa do vestibular entre os anos de 2005 e 2010. No período compreendido entre 2005 e 2009, o valor máximo da redação era de 20 pontos. Em 2010, o valor total passou a ser de 30 pontos e a diferença (discrepância) de nota permitida entre as correções era de 3 pontos. Não temos informação sobre o critério de discrepância para os anos anteriores. Outra limitação dos nossos dados foi que, nos casos de discrepância, não tivemos acesso à nota do terceiro corretor.

A literatura descreve dois métodos geralmente utilizados para a correção de redações em larga escala: o holístico e o analítico. A escolha adequada de um ou outro método depende de muitos fatores como, por exemplo, dos objetivos estabelecidos. Se o objetivo principal é a seleção dos estudantes, o método holístico pode ser suficiente. Entre os testes em larga escala que empregam esse método, podemos citar o TOEFL (*Test of English as a Foreign Language*). O método holístico consiste basicamente em atribuir um escore único com todos os critérios da avaliação considerados conjuntamente (exemplo: tema, coesão e ortografia). Já no método analítico, as redações não recebem apenas uma nota geral. Os avaliadores atribuem notas em diversos critérios, que combinados

geram um escore final. Um exemplo de avaliação em larga escala que emprega esse método é o Exame Nacional do Ensino Médio (Enem). Nesse exame, os examinadores avaliam a redação de acordo com cinco critérios (competências). Cada uma das competências recebe uma nota entre 0 e 200 pontos, podendo atingir um escore máximo de 1.000 pontos. No entanto, destacamos que o método analítico não tem sido muito utilizado para a correção de redações em larga escala por diversos fatores, como o alto custo (BACHA, 2001). Entretanto, algumas pesquisas indicam que, caso o objetivo principal seja mais educacional (f*eedback* para o aluno dos seus pontos fortes e fracos), o método analítico pode ser considerado mais apropriado (HAMP-LYONS, 1991), o que pode ser visto nas redações da segunda etapa da universidade pública que analisamos, já que se enquadram no método de correção analítico. O Quadro 2 mostra com detalhes os critérios empregados na correção.

QUADRO 2 – Critérios de correção das redações de 2010

FATORES AVALIADOS	PONTUAÇÃO MÁXIMA	CORREÇÃO DO EXAMINADOR
1. Adequação à proposta	6	
Tema	2	
Gênero / tipo textual	2	
Condições de circulação do texto	2	
2. Estratégias de textualização	13	
Coesão / coerência	4	
Progressão / paragrafação	3	
Informatividade	3	
Argumentação	3	
3. Aspectos Morfossintáticos	8	
Construção do período/ pontuação	4	
Concordância / regência / colocação	4	
4. Ortografia	3	
Total	30	

Fonte: Elaboração do autor.

Como indicado no Quadro 2, foi utilizado na correção das redações de 2010 um método de correção analítico com quatro critérios: adequação à proposta (máximo de 6 pontos), estratégias de textualização (máximo de 13 pontos), aspectos morfossintáticos (máximo de 8 pontos) e ortografia (máximo de 3 pontos). A nota final do candidato podia atingir 30 pontos, como resultado da soma de pontos obtidos nos quatro critérios de avaliação da redação.

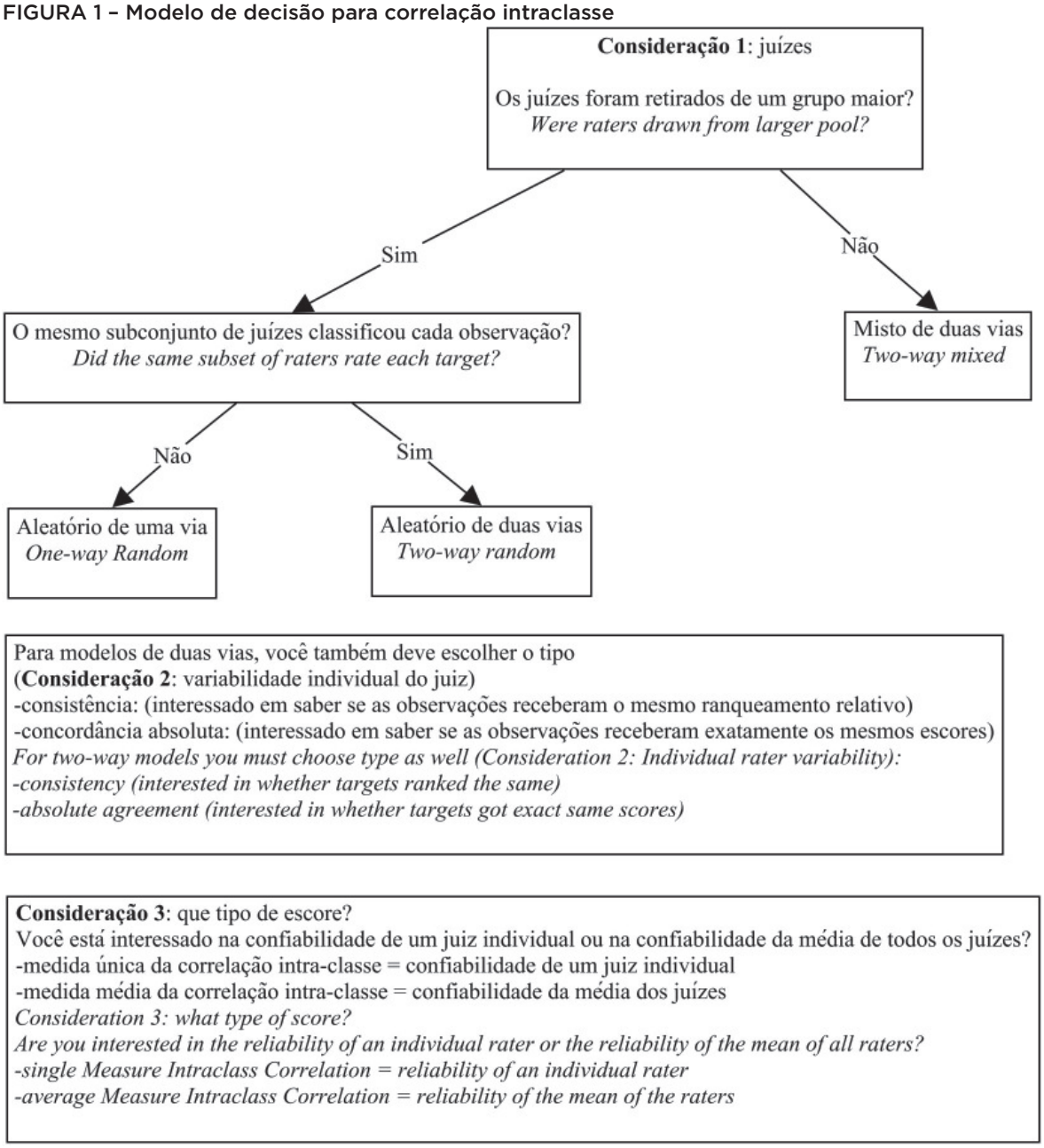
Destacamos que para os anos entre 2005 e 2009 não tivemos acesso aos critérios de correção (só obtivemos as notas totais de cada corretor). Além disso, não foi possível identificar os corretores pelos nomes. Dessa forma, a falta de informações para o referido período limitou a possibilidade de trabalho com os dados.

Como a nota das redações é uma variável contínua, utilizamos, nesta pesquisa, o coeficiente de correlação intraclasse – CCI, e, para realizar o cálculo das técnicas do nível de concordância entre juízes, adotamos o *software SPSS 20*. Para os anos de 2005 a 2010, produzimos uma análise geral da confiabilidade dos juízes considerando o processo seletivo como um todo. Além disso, no ano de 2010, fizemos também análises mais detalhadas do processo de correção para explicitar a gama variada de aplicações que as técnicas de concordância entre juízes possuem.

COEFICIENTE DE CORRELAÇÃO INTRACLASSE

Já afirmamos anteriormente que existem tipos diferentes de coeficiente de correlação intraclasse (CCI) e que ele pode ser estimado de diversas formas. Nesse sentido, indicamos uma série de questões que precisam ser respondidas para a escolha correta da forma de estimação do CCI (FIgura 1).

FIGURA 1 – Modelo de decisão para correlação intraclasse



Fonte: Romberg (2009, tradução nossa).

Como apontado na Figura 1, diversas questões devem ser respondidas para a utilização correta do CCI. Primeiramente, com relação aos juízes (consideração 1). Por exemplo: o mesmo subconjunto de juízes classificou cada observação? Em segundo lugar, para modelos de duas vias, a variabilidade individual do juiz está em jogo (consideração 2); em outras palavras, devemos estabelecer se nosso interesse está na

confiabilidade (consistência) ou na concordância (concordância absoluta) dos juízes. Por fim, determinamos o tipo de escore de nosso interesse: confiabilidade de um juiz individual ou confiabilidade da média de todos os juízes (consideração 3).

Destacamos que nossos resultados serão apresentados seguindo esse modelo de decisão.

RESULTADOS

REVISÃO DA LITERATURA

No total, encontramos na literatura brasileira 62 estudos que abordam as técnicas de verificação do nível de concordância entre juízes (25 dissertações/teses e 37 artigos). Nossa hipótese de que existia uma lacuna nas pesquisas educacionais brasileiras foi confirmada, pois 52 estudos foram realizados na área das ciências da saúde e apenas 10 nas ciências humanas (6 eram trabalhos relacionados à psicologia e apenas 4 se referiam a aplicações na educação). Santos *et al.* (2010), por exemplo, investigaram a confiabilidade dos julgamentos dos examinadores de prova escrita do processo seletivo do mestrado em Educação de uma universidade privada do Centro-Oeste. No total, foram 10 processos de seleção, entre os anos de 1994 e 2006. As atribuições de notas dos examinadores (juízes) foram analisadas ano a ano por meio do coeficiente de correlação de Pearson. Os resultados foram muito irregulares, com índices de correlação variando entre baixos (r = 0,15) e elevados (r = 0,89). Os autores finalizaram o estudo sugerindo procedimentos adicionais para melhorar a confiabilidade na avaliação das provas escritas do processo seletivo do mestrado.

Ainda com relação a nossa revisão da literatura, as estratégias de verificação do nível de concordância entre juízes utilizadas com maior frequência nos estudos foram: coeficiente *kappa* (23), coeficiente alfa de Cronbach (18), correlação de Pearson (12), coeficiente de correlação intraclasse (11), correlação de Spearman (7) e *kappa* ponderado (6). Outras técnicas foram empregadas com menor frequência, sendo alguns exemplos: *Kappa* múltiplo – *Fleiss' generalized kappa* (2), porcentagem de concordância absoluta (2), Teoria

de Resposta ao Item – modelo de Rasch (1) e Coeficiente Alfa de Krippendorff (1). Destacamos aqui dois aspectos: alguns estudos utilizam mais de uma técnica; e, apesar de a literatura apontar como método mais adequado para estimar a concordância entre juízes o coeficiente de correlação intraclasse, diversos trabalhos empregam outros tipos de correlação. Ainda assim, se somarmos todas as pesquisas que utilizaram correlação, essa passa a ser a técnica mais frequente.

CORREÇÕES DAS REDAÇÕES

Para os resultados das correções das redações, organizamos nossos dados a partir de diferentes perguntas de pesquisa. Destacamos que cada uma dessas perguntas exige uma aplicação diferente das técnicas de concordância entre juízes. Além disso, nas análises de todas as perguntas, excluímos as redações que tiveram nota zero, pois acreditamos que isso aumentaria artificialmente a concordância entre os juízes. Os critérios para o aluno receber nota zero eram muito claros (exemplo: entregar a redação em branco). De qualquer forma, o número de redações com nota zero no banco de dados foi pequeno.

PERGUNTA 1

Consideradas como um todo, qual foi a qualidade das correções das redações para o período entre 2005 e 2010? Em outras palavras: para fins de processo seletivo (vestibular), existiu uma confiabilidade mínima entre os juízes?

Consideração 1: juízes

Os juízes foram retirados de um grupo maior? Sim, dentre um grupo grande de corretores, somente dois eram designados para corrigir cada redação.

O mesmo subconjunto de juízes classificou cada observação? Não. Duplas diferentes de juízes corrigiram cada redação.

Isso implica um modelo aleatório de uma via. Portanto, não podemos realizar a escolha da consideração 2. Nesse caso, temos o resultado apenas da consistência (confiabilidade).

Consideração 3: Que tipo de escore?

Estamos interessados na confiabilidade da média de todos os juízes. Portanto, escolhemos a medida média da correlação intraclasse.

A Tabela 1 exibe, para todos os anos, o número de redações corrigidas, a menor e a maior nota, a média, o desvio padrão e a correlação intraclasse com o respectivo intervalo de confiança.

TABELA 1 – Análise das notas totais das redações no período entre 2005 e 2010

ANO	N	MIN.	MAX.	MÉDIA	DESVIO PADRÃO	CORRELAÇÃO INTRACLASSE (MÉDIA)	INTERVALO DE CONFIANÇA	
							LIMITE INFERIOR	LIMITE SUPERIOR
2005	4.266	1,5	20,0	11,8	3,1	0,86	0,858	0,874
2006	4.335	3,3	20,0	12,2	2,9	0,89	0,883	0,896
2007	2.324	1,0	18,5	11,6	2,5	0,87	0,864	0,885
2008	6.857	2,0	20,0	12,2	2,6	0,92	0,924	0,930
2009	10.223	1,4	20,0	11,6	2,7	0,91	0,911	0,918
2010	10.108	2,7	29,7	16,2	4,3	0,95	0,953	0,957

Fonte: Elaboração do autor.

Nota: Modelo aleatório de uma via. Intervalo de confiança 95 %.

A partir da Tabela 1, percebemos que um número grande de redações era corrigido todos os anos, sendo 2009 o ano em que mais redações foram corrigidas (10.223). A nota média das redações e a variabilidade das notas possuem valores próximos entre os anos de 2005 e 2009. Já no ano de 2010, a nota média das redações torna-se maior pelo fato de o valor total da redação ter sido alterado para 30 pontos, e, da mesma forma, a variabilidade dos dados também se torna maior. Quanto aos coeficientes de correlação intraclasse, eles variaram entre 0,86 e 0,95. Como essa análise se enquadra em uma decisão de alto impacto (aprovar pessoas em um vestibular de uma universidade pública), acreditamos que um critério de corte rigoroso precisa ser adotado. Nesse sentido, tendo como referência critérios para a tomada de decisões importantes sobre pessoas (valor mínimo de 0,8 a 0,9) (HAYS;

REVICKI, 2005), podemos afirmar que a confiabilidade da média de todos os juízes quanto à nota total das redações foi satisfatória. Também chama atenção o fato de o menor valor da correlação intraclasse ser de 2005 e o maior valor, de 2010. Isso provavelmente está relacionado à manutenção de membros da equipe de examinadores que, com o decorrer do tempo, aperfeiçoaram suas habilidades de correção.

Os resultados da pergunta 1 são importantes, mas apresentam algumas limitações, pois temos apenas dados sobre a confiabilidade (consistência) média de todos os juízes para a nota total da redação. Lembrando-se de que na nossa definição apresentada anteriormente, a confiabilidade entre juízes é uma medida da consistência entre avaliadores na ordenação ou posição relativa de avaliações de desempenho, independentemente do valor absoluto da classificação de cada avaliador. Assim, quando temos diferentes subconjuntos de juízes classificando cada participante, não podemos particionar a variância devido a juízes individuais. Portanto, com esses dados, não conseguimos identificar, por exemplo, membros da equipe que estejam com um padrão de confiabilidade abaixo do mínimo esperado. Isso fica diluído nesse tipo de análise geral. Além disso, também não podemos obter resultados sobre a concordância entre juízes (medida da consistência entre o valor absoluto das classificações dos avaliadores).

PERGUNTA 2

Analisando uma mesma dupla de juízes classificando as mesmas redações, qual a confiabilidade e a concordância da média dos dois examinadores?

Consideração 1: juízes

Os juízes foram retirados de um grupo maior? Sim, dentre um grupo grande de corretores, somente dois foram selecionados.

O mesmo subconjunto de juízes classificou cada observação? Sim. A mesma dupla de juízes corrigiu cada redação.

Isso implica um modelo aleatório de duas vias.

Consideração 2: variabilidade individual do juiz

Estamos interessados tanto na consistência (confiabilidade) quanto na concordância absoluta.

Consideração 3: Que tipo de escore?

Estamos interessados na confiabilidade da média dos juízes. Portanto, escolhemos a medida média da correlação intraclasse.

A Tabela 2 apresenta a confiabilidade e concordância média de juízes em 2010, tanto para a nota total da redação quanto para os quatro critérios de correção separadamente.

TABELA 2 – Confiabilidade e concordância média de juízes em 2010

JUÍZES	COEFICIENTES DE CORRELAÇÃO		CRITÉRIOS DE CORREÇÃO DAS REDAÇÕES				
			NOTA TOTAL	ADEQUAÇÃO À PROPOSTA (1)	ESTRATÉGIAS DE TEXTUALIZAÇÃO (2)	ASPECTOS MORFOSSINTÁTICOS (3)	ORTOGRAFIA (4)
Dalila e Eloisa	Alfa de Cronbach		0,93	0,81	0,75	0,68	0,77
	CCI	Consistência	0,93	0,81	0,75	0,68	0,77
		Concordância	0,86	0,81	0,70	0,60	0,77
Elisangela e Karine	Alfa de Cronbach		0,88	0,80	0,78	0,86	0,90
	CCI	Consistência	0,88	0,80	0,78	0,86	0,90
		Concordância	0,88	0,80	0,78	0,86	0,90

Fonte: Elaboração do autor.

Nota: Modelo aleatório de duas vias. Dalila e Eloisa, N= 374; Elisangela e Karine, N= 107 (nomes fictícios).

CCI: Coeficiente de correlação intraclasse.

Como já destacamos, a concordância mede com que frequência dois ou mais avaliadores atribuem exatamente a mesma classificação. A confiabilidade mede a semelhança relativa entre dois ou mais conjuntos de classificações. Nesse sentido, a Tabela 2 ilustra a diferença entre confiabilidade e concordância. No caso dos coeficientes de correlação intraclasse dos juízes Dalila e Eloisa, a concordância foi menor do que a confiabilidade para a nota total, o critério 2 e o critério 3. Como geralmente a concordância entre juízes é mais importante na educação quando tomamos decisões de alto impacto, podemos realizar uma série de considerações. Quanto à nota total, tomando novamente como critério de corte um valor

mínimo de 0,8 ou 0,9 (HAYS; REVICKI, 2005), a concordância pode ser considerada satisfatória. Além disso, como o método de correção da redação é analítico, é importante analisar também os quatro critérios separadamente. Assim, para a confiabilidade, os coeficientes variaram entre 0,68 e 0,81 (somente o critério 1 está acima do ponto de corte de 0,8). Já para a concordância (resultado mais importante), os coeficientes variaram entre 0,60 e 0,81 (os valores diminuem ainda mais nos critérios 2 e 3).

A dupla Elisangela e Karine, quando comparada à dupla Dalila e Eloisa, apresenta resultados mais confiáveis. Isso porque os coeficientes de confiabilidade e concordância são altos e possuem os mesmos valores. Quanto à nota total, apesar de indicar menor confiabilidade (0,88) do que Dalila e Eloisa (0,93), a dupla apresentou maior concordância (0,88). Isso se reflete também nos critérios da correção analítica, já que os coeficientes variaram entre 0,78 e 0,90, e foram, no geral, mais altos do que o valores da dupla Dalila e Eloisa, sendo que somente um dos critérios está abaixo do critério de corte de 0,8.

As análises dos quatro critérios separadamente são importantes por um motivo: geralmente, no método analítico, é recomendado um terceiro corretor tanto no caso de uma discrepância na nota total quanto de discrepância nos critérios adotados. O Enem, por exemplo, considera discordância entre avaliadores 200 pontos de discrepância na nota total ou 80 pontos em cada uma das cinco competências. Assim, a partir da Tabela 2, percebemos que os critérios 2 e 3 provavelmente foram problemáticos no caso da dupla Dalila e Eloisa. No entanto, como a universidade só trabalhava com a tolerância de até 3 pontos de diferença na nota total, esses fatores não foram considerados. Nesse caso, poderia ser solicitado um terceiro avaliador para esses critérios. Interessante destacar aqui que dados como esses podem ser utilizados, por exemplo, para identificar membros da equipe que estejam com um padrão de concordância abaixo do mínimo esperado. Isso pode resultar em um melhor treinamento da equipe.

Por fim, como já apontamos previamente, existem tipos diferentes de CCI e ele pode ser estimado de diversas formas.

O Coeficiente Alfa de Cronbach é muito citado na literatura. Por ser uma medida de consistência interna (confiabilidade), seu valor sempre é coincidente com o valor do coeficiente de correlação intraclasse de consistência/confiabilidade (Tabela 2).

PERGUNTA 3

Analisando uma mesma dupla de juízes classificando as mesmas redações, qual a confiabilidade de qualquer examinador individual? Eles estão utilizando pontos de ancoragem diferentes em suas classificações?

Consideração 1: juízes

Os juízes foram retirados de um grupo maior? Sim, dentre um grupo grande de corretores, somente dois foram selecionados.

O mesmo subconjunto de juízes classificou cada observação? Sim. A mesma dupla de juízes corrigiu cada redação.

Isso implica um modelo aleatório de duas vias.

Consideração 2: variabilidade individual do juiz

Estamos interessados apenas na concordância absoluta, pois queremos verificar se os juízes estão usando pontos de ancoragem diferentes em suas classificações.

Consideração 3: Que tipo de escore?

Estamos interessados na concordância de um juiz individual. Portanto, escolhemos a medida única (*single measures*) da correlação intraclasse.

A Tabela 3 mostra a concordância individual e média de juízes em 2010, tanto para a nota total da redação quanto para os quatro critérios de correção separadamente.

Em primeiro lugar, apesar de a pergunta 3 necessitar apenas da medida única da correlação intraclasse , também incluímos na Tabela 3 a média das correlações por um motivo: ilustrar a importância de utilizar múltiplos juízes em avaliações de alto impacto. Dessa forma, o CCI individual e o CCI médio indicam a segurança que temos no processo avaliativo se estivermos baseados em apenas um juiz ou na

média dos juízes. Obviamente, para fins de seleção, utilizaremos sempre o CCI médio. No entanto, o CCI individual nos fornece outro tipo de informações igualmente importantes.

Dessa maneira, a análise dos resultados da dupla Felipe e Wallace (nomes fictícios) indica algumas questões interessantes. Primeiramente, apesar de a concordância média dos dois juízes estar em patamar aceitável para avaliações de alto impacto na nota total e no critério 1, o CCI individual indica que os juízes provavelmente estão usando pontos de ancoragem diferentes em alguns critérios de correção, como no critério 3, que parece ser o mais problemático, por apresentar uma concordância muito baixa. Por isso, tomados em conjunto, esses resultados mostram que uma concordância média satisfatória na nota total pode esconder alguns problemas, nesse caso, referentes à falta de concordância em critérios do método de correção analítico (mesma crítica apresentada na pergunta 2) e à utilização de pontos de ancoragem diferentes pelos juízes em alguns critérios.

Tabela 3 - Concordância individual e média de juízes em 2010

CRITÉRIOS DE CORREÇÃO DAS REDAÇÕES	INDIVIDUAL			MÉDIO		
	CCI	INTERVALO DE CONFIANÇA		CCI	INTERVALO DE CONFIANÇA	
		LIMITE INFERIOR	LIMITE SUPERIOR		LIMITE INFERIOR	LIMITE SUPERIOR
Nota total	0,76	0,550	0,864	0,86	0,710	0,927
Adequação à proposta (1)	0,70	0,635	0,758	0,82	0,777	0,863
Estratégias de textualização (2)	0,64	0,524	0,739	0,78	0,688	0,850
Aspectos morfossintáticos (3)	0,10	-0,017	0,216	0,18	-0,034	0,355
Ortografia (4)	0,63	0,554	0,699	0,77	0,713	0,823

Fonte: Elaboração do autor.

Nota: Modelo aleatório de duas vias. Intervalo de confiança 95 %. Felipe e Wallace. N= 265 (nomes fictícios).

CCI: Coeficiente de Correlação Intraclasse.

Por fim, a Tabela 3 também evidencia falta de precisão dos resultados, pois alguns intervalos de confiança são muito amplos. Esse erro de medida provavelmente está ligado à amostra pequena de redações analisadas (N= 265). Esse fato sugere-nos que, para avaliar a qualidade da correção da dupla de juízes, é

interessante ter o maior número possível de observações.

PERGUNTA 4

O critério de 3 pontos de discrepância na nota total da redação é suficiente?

Para responder essa questão, selecionamos aleatoriamente no banco de dados casos em que os juízes discordaram em mais de 3 pontos (N=276) na nota total das redações e casos nos quais a discordância dos juízes estava num intervalo entre 2 e 3 pontos (N=423). Na primeira situação (discordância em mais de 3 pontos), a confiabilidade média dos juízes foi de apenas 0,19. Na segunda situação, a confiabilidade média foi de 0,88. Lembramos aqui, mais uma vez, que, como temos duplas diferentes de juízes corrigindo cada redação, apenas a confiabilidade/consistência média é analisada (modelo aleatório de uma via).

Portanto, o critério de 3 pontos de discrepância na nota total da redação parece ser sensível o suficiente para indicar baixa confiabilidade entre os juízes e a necessidade de um terceiro corretor. No entanto, algumas limitações que apontamos em outras perguntas permanecem aqui, entre elas a impossibilidade de obter resultados sobre a concordância entre juízes e a falta de indicadores por parte da universidade para analisar a discrepância nos quatro critérios do método de correção analítica.

Tomadas em conjunto, as quatro perguntas de pesquisa discutidas anteriormente ilustram a complexidade da análise de um processo avaliativo. Dizer se os resultados de uma avaliação são confiáveis ou não, envolve muitas variáveis.

CONSIDERAÇÕES FINAIS

Encontramos na literatura brasileira 62 pesquisas que utilizaram estratégias de verificação da concordância entre juízes. Nossa hipótese de que existia uma lacuna nas pesquisas educacionais brasileiras foi confirmada, pois a maioria dos estudos foi realizada na área das ciências da saúde. Assim, recomendamos tanto uma maior utilização das técnicas de

concordância de juízes quanto à realização de mais pesquisas na área educacional.

Quanto à correção das redações, para os anos de 2005 a 2010, realizamos uma análise geral da confiabilidade dos juízes considerando o processo seletivo como um todo. No ano de 2010, fizemos também análises mais detalhadas do processo de correção para explicitar a gama variada de aplicações que as técnicas de concordância entre juízes possuem. A partir disso, percebemos que alguns resultados foram satisfatórios (exemplo: confiabilidade média dos juízes para as notas totais das redações) e outros insatisfatórios (exemplo: concordância baixa em alguns critérios de correção). Isso ilustra a complexidade da análise de um processo avaliativo, de modo que afirmar se os resultados de uma avaliação são confiáveis ou não envolve muitos fatores.

Com relação à concordância entre juízes, indicamos um ponto importante para pensar o critério de corte de todas as técnicas apresentadas anteriormente: de um modo geral, os pesquisadores afirmam que quanto maior as consequências resultantes da avaliação, maior a necessidade de uma concordância entre juízes alta (LEBRETON; SENTER, 2008). Portanto, não existe uma regra geral única, e devemos estar especialmente atentos para situações que envolvem decisões de alto impacto. Nesse sentido, as perguntas de pesquisa que usamos exemplificam essa questão.

Concluímos apontando uma questão central para os educadores hoje: qual o nível de confiabilidade e concordância dos diversos processos avaliativos que acontecem na área educacional? Se não conseguirmos responder a essa questão, não saberemos se as diversas avaliações estão atingindo o resultado esperado. Como já discutimos, a eficácia dos processos avaliativos está condicionada, dentre outros fatores, pelo nível de concordância entre juízes.

REFERÊNCIAS

ALTMAN, D. *Practical statistics for medical research. Boca Raton*, FL: CRC, 1991.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT

IN EDUCATION. *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association, 1999.

ANDRADE, M.; SHIRAKAWA, I. Versão brasileira do Defense Style Questionnaire (DSQ) de Michael Bond: problemas e soluções. *Revista de Psiquiatria do Rio Grande do Sul*, Porto Alegre, v. 28, n. 2, p. 144-160, 2006.

BACHA, N. Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, v. 29, p. 371-383, 2001.

BLAND, J. M.; ALTMAN, D. G. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput. Biol. Med.*, v. 20, n. 5, p. 337-340, 1990.

BLOOD, E.; SPRATT, K. F. *Disagreement on Agreement: Two Alternative Agreement Coefficients.* SAS Global Forum, 2007.

BRUSCATO, W. L.; IACOPONI, E. Validade e confiabilidade da versão brasileira de um inventário de avaliação de relações objetais. *Rev. Bras. Psiquiatr.*, São Paulo, v. 22, n. 4, p. 172-177, 2000.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v. 20, p. 37-46, 1960.

_____. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, v. 70, p. 213-220, 1968.

CROCKER, L.; ALGINA, J. *Introduction to Classical and Modern Test Theory.* Belmont, CA: Wadsworth Group, 2009.

DEL-BEN, C. M.; VILELA, J. A. A.; CRIPPA, J. A. S.; HALLAK, J. E. C.; LABATE, C. M.; ZUARDI, A. W. Confiabilidade da Entrevista Clínica Estruturada para o DSM-IV – Versão Clínica traduzida para o português. *Rev. Bras. Psiquiatr.,* São Paulo, v. 23, n. 3, p. 156-159, 2001.

EMBRETSON, S. E.; REISE, S. P. *Item response theory for psychologists*. New York: Routledge, 2000.

FLEISS, J. *Statistical methods for rates and proportions.* New York: John Wiley & Sons, 1981.

FONSECA, R.; SILVA, P.; SILVA, R. Acordo inter-juízes: O caso do coeficiente kappa. *Laboratório de Psicologia*, Lisboa, v. 5, n.1, p. 81-90, 2007.

FRAGA-MAIA, H.; SANTANA, V. S. Concordância de informações de adolescentes e suas mães em inquérito de saúde. *Revista de Saúde Pública*, São Paulo, v. 39, n. 3, p. 430-437, 2005.

GRAHAM, M.; MILANOWSKI, A.; MILLER, J. *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Research Report, 2012.

GWET, K. *Handbook of inter-rater reliability:* How to estimate the level of agreement between two or multiple raters. Gaithersburg: Stataxis, 2001.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. *Análise multivariada de dados*. 5. ed. Porto Alegre: Bookman, 2005. 593 p.

HAMP-LYONS, L. Scoring procedures for ESL contexts. In: HAMP-LYONS, L. (Ed.). *Assessing second language writing in academic contexts.* Norwood, NJ: Ablex, 1991. p. 241–276.

HANEY, W.; RUSSELL, M.; BEBELL, D. Drawing on education: using drawings to document schooling and support changes. *Harvard Educational Review*, v. 74, n. 3, 241-271, 2004.

HAYS, R. D.; REVIKI, D. A. Reliability and validity (including responsiveness). In: FAYERS, P. M.; HAYS, R. D. (Ed.). *Assessing quality of life in clinical trials:* Methods and practice. NY: Oxford University Press, 2005.

JORBA, J.; SANMARTÍ, N. A função pedagógica da avaliação. In: BALLESTER, M. (Org.). *Avaliação como apoio à aprendizagem.* Porto Alegre: Artmed, 2003. cap 2, p.23-45.

KING, J. E. Software Solutions for Obtaining a Kappa-Type Statistic for Use with Multiple Raters. In: ANNUAL MEETING OF THE SOUTHWEST EDUCATIONAL RESEARCH ASSOCIATION, 2004, Dallas, EUA. *Anais…* Dallas: 2004.

LANDIS, J. R.; KOCH, G. G. A one way components of variance model for categorical data. *Biometrics*, v. 33, p. 671–679, 1977.

LEBRETON, J. M.; SENTER, J. L. Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, v. 11, n. 4, p. 815-852, 2008.

LINACRE, J. M. Rating, judges and fairness. *Rasch Measurement Transactions*, v. 12, n. 2, p. 630-1, 1998.

LU, L.; SHARA, N. *Reliability analysis:* calculate and compare intra-class correlation coefficients (ICC) in SAS. NESUG, 2007.

PERROCA, M. G.; GAIDZINSKI, R. R. Instrumento de classificação de pacientes de perroca: teste de confiabilidade pela concordância entre avaliadores – correlação. *Rev. Esc. Enferm.*, São Paulo, v. 36, n. 3, p. 245-252, 2002.

_____. Avaliando a confiabilidade interavaliadores de um instrumento para classificação de pacientes – coeficiente Kappa. *Rev. Esc. Enferm.*, São Paulo, v. 37, n. 1, p. 72-80, 2003.

POLANCZYK, G. V.; EIZIRIK, M.; ARANOVICH, V.; DENARDIN, D.; SILVA, T. L.; CONCEIÇÃO, T. V.; PIANCA, T. G.; ROHDE, L. A. Interrater agreement for the schedule for affective disorders and schizophrenia epidemiological version for school-age children (K-SADS-E). *Rev. Bras. Psiquiatr.,* São Paulo, v. 25, n. 2, p. 87-90, 2003.

PRIMI, R.; MIGUEL, F. K.; COUTO, G.; MUNIZ, M. Precisão de avaliadores na avaliação da criatividade por meio da produção de metáforas. *Psico-USF*, Itatiba, v. 12, n. 2, p. 197-210, 2007.

QUINTANA, H. E. O portfólio como estratégia para a avaliação. In: BALLESTER, M. (Org.). *Avaliação como apoio à aprendizagem.* Porto Alegre: Artmed, 2003. cap 16, p.163-173.

ROMBERG, A. *Intraclass correlation coefficients.* Reliability and more. 2009. Disponível em: <http://www.docstoc.com/docs/112692917/Intraclass-correlation-coefficients>. Acesso em: 07 jan. 2012.

SANTOS, T. M. de B. M. dos; MONTEIRO, V. R. V.; JUNIOR, J. F. R. Confiabilidade dos julgamentos de avaliadores de prova escrita na seleção para o mestrado. *Est. Aval. Educ.*, São Paulo, v. 21, n. 46, p. 363-374, maio/ago. 2010.

SCHUSTER, C. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, v. 64, p. 243-253, 2004.

STEMLER, S. E. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, v. 9, n. 4, 2004.

STUFFLEBEAM, D. L. (Org.). *Educational evaluation & decision making*. Bloomington: Phi Delta Kappa, 1971.

TINSLEY, H. E. A.; WEISS, D. J. Interrater reliability and agreement. In: TINSLEY, H. E. A.; BROWN, S. D. (Ed.). *Handbook of applied multivariate statistics and mathematical modeling*. New York: Academic Press, 2000. p. 95-124.

URBINA, S. *Fundamentos da testagem psicológica*. Porto Alegre: Artmed, 2007.

VENTURA, M. M.; BOTTINO, C. M. C. Estudo de confiabilidade da versão em português de uma entrevista estruturada para o diagnóstico de demência. *Revista da Associação Médica Brasileira*, São Paulo, v. 47, n. 2, p. 110-116, 2001.

DANIEL ABUD SEABRA MATOS
Doutor em Educação pela Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil. Doutorado sanduíche na University of Florida, Estados Unidos. Professor do Departamento de Educação da Universidade Federal de Ouro Preto (Ufop), Ouro Preto, Minas Gerais, Brasil
danielmatos@ichs.ufop.br

Recebido em: MAIO 2014
Aprovado para publicação em: OUTUBRO 2014