ANÁLISE DE CORRELAÇÃO NA OBTENÇÃO DA CONFIABILIDADE E ÍNDICE DE DISCRIMINAÇÃO EM INSTRUMENTO DE AVALIAÇÃO DE ENSINO

Sachiko Araki Lira
Instituto Paranaense de Desenvolvimento Econômico e Social – Ipardes
Rua Máximo João Kopp, 274 – Santa Cândida – Bloco 2 - CEP 82.630-900
Curitiba - PR
sachiko@onda.com.br

Anselmo Chaves Neto
Universidade Federal do Paraná
Programa de Pós-Graduação em Métodos Numéricos em Engenharia
Centro Politécnico - CEP 81.531-990
Curitiba – PR
anselmo@est.ufpr.br

Resumo. *A avaliação de resultados é necessária em todas as áreas de atividade humana. Trata-se de uma fase importante não somente em processos de produção, mas também quando se deseja conhecer a satisfação dos clientes em relação aos produtos e serviços adquiridos. Em se tratando de ensino, o mercado de trabalho exige, cada vez mais, profissionais competentes, que possam atuar de forma eficiente e eficaz, tornando, assim, imprescindível a necessidade da avaliação, tanto da assimilação do conteúdo pelos alunos, quanto da qualidade do ensino pelos professores. Chaves Neto e Turim (2003) comparam o processo de ensino e avaliação com o ciclo de Shewhart, conhecido como ciclo PDCA (plan, do, check, act), significando planejar, fazer (executar), avaliar e realimentar. Deve-se planejar uma ação, aplicá-la, avaliar os resultados e realimentar o planejamento de forma contínua, buscando sempre o aperfeiçoamento contínuo do processo de ensino. As condições básicas de um bom instrumento de medida educacional, segundo Chaves Neto e Turim, são a validade, a confiabilidade (fidedignidade), a objetividade e a praticabilidade. Já as propriedades dos itens que compõem o teste são o grau de dificuldade e o índice de discriminação. A Teoria de Resposta ao Item (TRI) possui modelos que relacionam a habilidade do estudante* (θ), *o grau de dificuldade do item* (b) *e o índice de discriminação do item* (a). *Por meio desta teoria pode-se estimar os parâmetros do modelo e classificar tanto os itens quanto os testes. A confiabilidade do teste, bem como o grau de discriminação de cada item, podem ser estimados pelo coeficiente de correlação. O objetivo deste trabalho é apresentar os diferentes métodos para estimar a confiabilidade: teste-reteste, forma paralela, split-half, Spearman-Brown, Kuder-Richardson e Alfa de Cronbach.*

Palavras-chave: *correlação, confiabilidade.*

1. INTRODUÇÃO

A avaliação é necessária não somente em processos de produção, mas também quando se deseja conhecer a satisfação dos clientes em relação aos produtos e serviços adquiridos.

Em se tratando de ensino, o mercado de trabalho exige, cada vez mais, profissionais competentes, que possam atuar de forma eficiente e eficaz, tornando imprescindível, a necessidade da avaliação, tanto da assimilação do conteúdo pelos alunos, quanto da qualidade do ensino pelos professores.

Chaves Neto e Turim (2003) comparam o processo de ensino e avaliação com o ciclo de Shewhart, conhecido como ciclo PDCA (plan, do, check, act), significando planejar, fazer (executar), avaliar e realimentar. Deve-se planejar uma ação, aplicá-la, avaliar os resultados e realimentar o planejamento de forma contínua, buscando sempre o aperfeiçoamento contínuo do processo de ensino.

As condições básicas de um bom instrumento de medida educacional, segundo Chaves Neto e Turim (2003), são a validade, a confiabilidade (fidedignidade), a objetividade e a praticabilidade. Já as propriedades dos itens que compõem o teste são o grau de dificuldade e o índice de discriminação.

A Teoria de Resposta ao Item (TRI) possui modelos que relacionam a habilidade do estudante (θ), o grau de dificuldade do item (b) e o índice de discriminação do item (a). Por meio desta teoria pode-se estimar os parâmetros do modelo e classificar tanto os itens quanto os testes.

Já, na Teoria Clássica de Avaliação, o grau de dificuldade de um item é representado pela percentagem de acertos do total de examinandos a ele submetido e a confiabilidade do teste, bem como o grau de discriminação de cada item, podem ser estimados pelo coeficiente de correlação.

O índice de discriminação de um item pode ser obtido através do cálculo do coeficiente de correlação entre os pares (x_j, y_j), onde $j = 1, 2, ..., n$ representa o número do examinando; y_j, o escore do teste do examinando j; e x_j é uma variável aleatória dicotômica, assumindo zero ou 1, conforme o examinando acerte ou erre o item.

O presente trabalho está dividido em cinco seções, incluindo esta. Na segunda seção são apresentados os métodos de análise de correlação utilizados para a análise dos índices de discriminação e confiabilidade. Na terceira seção discute-se o índice de discriminação utilizando o coeficiente de correlação, e, na quarta, mostram-se os métodos para a obtenção do índice de confiabilidade e finalmente a conclusão.

2. MÉTODOS DE ANÁLISE DE CORRELAÇÃO

São discutidos, neste trabalho, os métodos de Análise de Correlação utilizados para estimar a confiabilidade do instrumento de avaliação (teste) e o índice de discriminação de itens que compõe o teste.

O Coeficiente de Correlação Linear de Pearson mede a correlação entre duas variáveis contínuas, tendo, como suposição básica, a relação linear entre elas. É conhecido também como Coeficiente de Correlação do Momento Produto. Este coeficiente, ρ, é um valor real situado no intervalo $[-1, 1]$.

Quanto mais próximo de 1 for o valor de $|\rho|$, maior será a relação entre as variáveis envolvidas na análise. Segundo Ferguson (1981) e Chaves Neto e Turim (2003), o estimador do Coeficiente Linear de Pearson entre as variáveis X e Y tem a forma apresentada seguir:

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}} \qquad (1)$$

onde:
$\overline{X}$ é a média amostral da variável X ;
$\overline{Y}$ é a média amostral da variável Y.

Já o Coeficiente de Correlação Ponto Bisserial é indicado quando uma das variáveis é contínua e a outra é dicotômica (Guilford, 1950; McNemar, 1969 e Ferguson, 1981). Quando se calcula este coeficiente a partir de seus dados originais, obtém-se o mesmo coeficiente que o Coeficiente de Correlação Linear de Pearson. O Coeficiente de Correlação Ponto Bisserial é estimado através de:

$$\hat{\rho}_{p_b} = \frac{\overline{X}_t - \overline{X}_p}{S_t}\sqrt{\frac{p}{1-p}} \qquad (2)$$

onde:
$\overline{X}_t$ é a média dos escores;
$\overline{X}_p$ é a média dos escores dos examinandos que responderam ao item corretamente;
p é a proporção de examinandos que responderam ao item corretamente.

Quando o interesse é medir a correlação entre duas variáveis dicotômicas, ou seja, do tipo certo ou errado, sucesso ou insucesso, conforme ou não conforme, pode-se utilizar o Coeficiente de Correlação Phi (Guilford, 1950; McNemar, 1969 e Ferguson, 1981). Este coeficiente é utilizado no método de Kuder-Richardson, que será apresentado adiante, para estimar a confiabilidade de instrumentos que envolvem itens que assumem resultados tipo zero e 1. O estimador do Coeficiente de Correlação Phi é apresentado a seguir:

$$\hat{\phi} = \frac{(ad - bc)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \qquad (3)$$

onde:
a, b, c, d são as freqüências em uma tabela de contingência 2x2.

Ainda, na avaliação da discriminação de um item é usado o Coeficiente de Correlação Bisserial. Este é indicado quando se tem duas variáveis contínuas, no entanto uma delas pode ser dicotomizada (Guilford, 1950 e McNemar, 1969). Por exemplo, os examinandos podem ser separados em duas categorias, "aprovados" e "reprovados", de acordo com o escore obtido. Este coeficiente é estimado através de:

$$\hat{\rho}_b = \frac{\overline{X}_p - \overline{X}_t}{S_t}\frac{p}{y} \quad \text{ou} \quad \hat{\rho}_b = \frac{\overline{X}_p - \overline{X}_q}{S_t}\frac{p \times q}{y} \qquad (4)$$

onde:
$\overline{X}_p$ é a média dos valores de X para o grupo superior (grupo cujos valores de X estão acima do ponto de dicotomização da variável Y);
$\overline{X}_q$ é a média dos valores de X para o grupo inferior (grupo cujos valores de X estão abaixo do ponto de dicotomização da variável Y);
$\overline{X}_t$ é a média da variável X;
p é a proporção de casos do grupo superior (grupo cujos valores de X estão acima do ponto de dicotomização da variável Y);
q é a proporção de casos do grupo inferior (grupo cujos valores de X estão abaixo do ponto de dicotomização da variável Y);

y é a ordenada da distribuição normal no ponto de dicotomização (p) da variável Y.

Inicialmente obtém-se o valor de z, correspondente à área menor ou igual a p e calcula-se a ordenada correspondente na f.d.p., $y = f(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$.

3. ANÁLISE DE DISCRIMINAÇÃO DE ITEM

Discriminação é a capacidade de um item diferenciar examinandos que possuem habilidades distintas, muitas vezes muito próximas. Um item muito fácil pode não fornecer um índice de discriminação desejável, pois a grande maioria dos examinandos poderá acertá-lo, o mesmo acontecendo com um item muito difícil, em que muitos examinandos poderão errar, conseqüentemente um item facílimo ou dificílimo não tem a propriedade da discriminação, que é muito desejável nos instrumentos de avaliação.

O índice de discriminação do item pode ser obtido calculando-se o coeficiente de correlação entre o escore total (Y) e a resposta do examinando ao item (X), que assume valor 1, se respondeu corretamente, e zero caso contrário. É possível utilizar o Coeficiente de Correlação Linear de Pearson ou o Coeficiente de Correlação Ponto Bisserial e o Coeficiente de Correlação Bisserial (Chaves Neto e Turim, 2003).

Quanto maior o coeficiente de correlação, maior a discriminação do item, indicando que a relação entre o escore total e o item respondido corretamente é grande.

Considere-se uma situação onde se tem um teste de múltipla escolha composto por cinco itens, em que a resposta é única, tendo sido atribuído o valor 1 se a resposta está correta e zero se errada, aplicado a vinte examinandos. Os dados apresentados na Tab. (1) foram obtidos pelo processo de simulação. As estatísticas dos itens se encontram na Tab. (2).

Tabela 1. Resultados das respostas aos itens obtidos pelo processo de simulação

Itens	Examinandos																				Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	19
2	1	1	1	0	1	1	1	0	1	0	0	1	0	0	1	1	1	1	1	1	14
3	1	0	1	0	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	9
4	1	1	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	9
5	1	1	1	0	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	8
Total	5	4	5	1	5	5	5	2	5	2	1	5	1	1	2	2	2	2	2	2	59

Tabela 2. Número de acertos, total de examinandos e coeficiente de correlação linear de Pearson segundo itens

Itens	Número de acertos	Total de examinandos	Correlação linear de Pearson
1	19	20	0,13
2	14	20	0,65
3	9	20	0,77
4	9	20	0,89
5	8	20	0,97

Analisando-se a Tab.(2) é possível concluir que o item 5 é o de maior discriminação, por apresentar o maior coeficiente de correlação.

4. ANÁLISE DE CONFIABILIDADE

Entende-se por confiabilidade, em educação, a consistência dos escores obtidos pelos examinandos em determinado teste. Um instrumento é confiável se produz resultados estáveis quando aplicado em diferentes momentos. Esta estabilidade significa a confiabilidade do instrumento. Mede-se a confiabilidade através da Análise de Correlação. O índice de confiabilidade varia entre 0 e 1. Quanto mais próximo de 1 for o índice, mais confiável é o instrumento de avaliação.

Diferentes métodos para estimar a confiabilidade foram desenvolvidos com base na teoria da Análise de Correlação, apresentados em Ferguson (1981).

4.1. Método do Teste-Reteste

Neste método, o mesmo instrumento de medida (teste) é aplicado em duas ocasiões distintas para o mesmo grupo de examinandos. Calcula-se, então, o Coeficiente de Correlação Linear de Pearson para o conjunto dos escores obtidos pelos examinandos nos dois testes.

Existem alguns fatores que afetam a efetividade desta técnica, como o tempo decorrido entre a aplicação dos testes. Quanto maior o tempo transcorrido entre os dois testes menor é a correlação. Este teste é freqüentemente utilizado para calcular a confiabilidade de testes escritos, sendo conhecido também como coeficiente de estabilidade.

4.2. Método da Forma Paralela

Este método é também conhecido como forma equivalente. Nele, aplica-se um teste da forma "A" para grupo de examinandos, e imediatamente após um teste da forma "B", para o mesmo grupo, com o mesmo conteúdo. As duas formas são feitas com os mesmos tipos de itens (perguntas). O Coeficiente de Correlação Linear de Pearson é calculado para o conjunto de escores obtidos nos testes das formas "A" e "B".

4.3. Método *Split-Half*

A vantagem deste método é que necessita somente de um conjunto de dados. Considera-se, normalmente, o número de acertos das questões pares e o número de acertos das questões ímpares. Ou, ainda, as duas primeiras questões para o primeiro escore, as próximas duas para o segundo escore, e assim alternadamente. Não é aconselhável fazer a divisão dos itens exatamente ao meio, pois é comum as primeiras questões serem mais fáceis do que as últimas. O Coeficiente de Correlação Linear de Pearson é calculado para o conjunto de escores, pares e ímpares.

4.4. Método de Spearman-Brown

O método de Spearman-Brown permite calcular o Coeficiente de Confiabilidade do teste com um número maior de itens, conhecendo-se a confiabilidade de um número menor de itens. Obtido o Coeficiente de Confiabilidade pelo método Split-Half, através do método de Spearman-Brown é possível estimar o coeficiente para o total, por meio da fórmula:

$$\hat{\rho}_{X,X} = \frac{2\hat{\rho}_{X_1,X_2}}{1+\hat{\rho}_{X_1,X_2}} \tag{5}$$

onde:

$\hat{\rho}_{X,X}$ é a confiabilidade estimada para o total;

$\hat{\rho}_{X_1,X_2}$ é a confiabilidade para a metade do teste.

Foram desenvolvidos métodos que utilizam estatísticas de itens, conhecidos como de consistência interna, e que são os mais utilizados.

4.5. Método de Kuder-Richardson

$$\hat{\rho}_{X,X} = \frac{n}{n-1}\frac{S_X^2 - \sum_{i=1}^{n} p_i q_i}{S_X^2} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} p_i q_i}{S_X^2}\right] \tag{6}$$

onde:

n é o número de itens;

S_X^2 é a variância de escores do teste;

$\sum_{i=1}^{n} p_i q_i$ é a soma do produto de proporções de acertos e erros em cada item.

4.6. Método de Alfa de Cronbach

Lee Cronbach generalizou a expressão de Kuder-Richardson para o caso onde os itens não são todos dicotômicos (Cronbach, 1951). Esta expressão recebeu o nome de "Alfa de Cronbach", apresentada a seguir:

$$\alpha = \frac{n}{n-1}\frac{S^2 - \sum_{i=1}^{n} S_i^2}{S^2} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} S_i^2}{S^2}\right] \tag{7}$$

onde:

n é o número de itens, S^2 é a variância de escores do teste e $\sum_{i=1}^{n} S_i^2$ é a soma das variâncias dos escores no item i.

Os quatro primeiros métodos apresentados permitem medir a confiabilidade considerando-se os escores totais; já os dois últimos permitem medir os escores totais e os itens que os compõem. Como um exemplo prático, a Fig. (1) abaixo, mostra o índice de confiabilidade do teste, utilizando o Coeficiente de Correlação Linear de Pearson entre os escores obtidos em duas avaliações, empregando o mesmo instrumento. Os escores apresentados na Tab. (3) foram obtidos pelo processo de simulação.

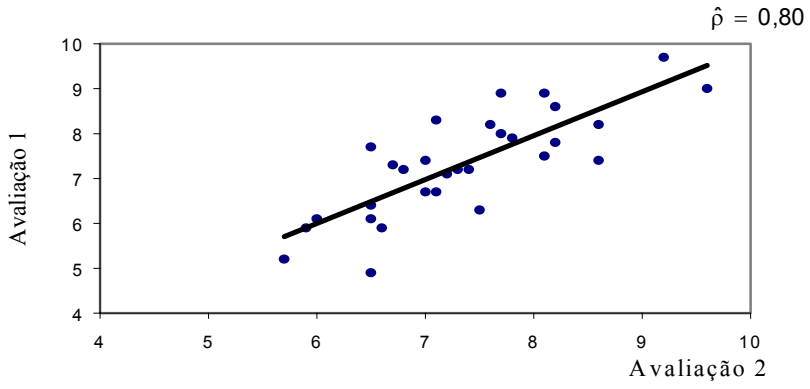

Figura 1. Índice de confiabilidade do teste

Tabela 3. Escores obtidos pelos examinandos nas avaliações 1 e 2

Avaliação	Examinandos														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	6,5	7,4	8,2	7,0	8,1	7,2	8,6	6,0	6,7	9,6	8,2	8,1	6,8	6,5	7,1
2	7,7	7,2	7,8	6,7	8,9	7,1	7,4	6,1	7,3	9,0	8,6	7,5	7,2	6,4	8,9
Avaliação	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	6,5	7,4	8,2	7,0	8,1	7,2	8,6	6,0	6,7	9,6	8,2	8,1	6,8	6,5	7,1
2	7,7	7,2	7,8	6,7	8,9	7,1	7,4	6,1	7,3	9,0	8,6	7,5	7,2	6,4	8,9

Quando se tem testes compostos por questões de múltipla escolha, em que se atribuem o valor 1 para respostas corretas, e o valor zero para as incorretas, o índice de confiabilidade é obtido pelo Método de Kuder-Richardson, como já mencionado anteriormente. Considerando a mesma situação apresentada na Tab. (1), tem-se as seguintes estatísticas:

Tabela 4. Proporção de acertos e erros segundo itens

Itens	Proporção de acertos (p)	Proporção de erros (q=1-p)	pxq
1	0,95	0,05	0,0475
2	0,70	0,30	0,2100
3	0,45	0,55	0,2475
4	0,45	0,55	0,2475
5	0,40	0,60	0,2400
Total	-	-	0,9925

O desvio padrão do escore total do teste dos examinandos foi de $S_x = 1{,}6694$. Tem-se, então, que: $\hat{\rho}_{xx} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} p_i q_i}{S_x^2}\right] = \frac{5}{4}\left[1 - \frac{0{,}9925}{(1{,}6694^2)}\right] = 0{,}8048$

Este índice permite afirmar que o instrumento é confiável, pois quanto mais próximo de 1 for o índice, melhor é o instrumento de avaliação.

Em situações em que o teste é composto por itens que recebem diferentes pontuações, o índice de confiabilidade é estimado pelo Método de Alfa de Cronbach. Cita-se, a seguir, um exemplo ilustrativo. Os dados da Tab. (5) foram obtidos pelo processo de simulação.

Tabela 5. Resultado das respostas segundo itens

Itens	Examinandos																				Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	3	2	3	1	3	1	3	2	3	1	2	3	1	1	2	2	2	2	2	2	41
2	1	1	1	0	1	1	1	0	1	0	1	1	0	0	1	1	1	1	1	1	15
3	3	2	3	1	3	3	1	1	1	1	1	3	1	1	1	2	1	1	1	2	33
4	1	1	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	09
5	3	3	3	1	3	2	1	1	1	1	2	2	1	1	1	1	1	1	1	1	31
Total	11	9	11	3	11	8	7	4	7	4	6	10	3	3	5	6	5	5	5	6	129

As estatísticas apresentadas na Tab. (6) foram obtidas a partir da Tab.(5).

Tabela 6. Desvios padrão e variâncias segundo Itens

Itens	Desvio Padrão	Variância
1	0,7592	0,5763
2	0,4443	0,1974
3	0,8751	0,7658
4	0,5104	0,2605
5	0,8256	0,6816
Total	-	2,4816

Tem-se que o desvio padrão dos escores é $S = 2{,}7237$; portanto, o índice de confiabilidade é $\alpha = 0{,}8319$, indicando que o instrumento é confiável.

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} S_i^2}{S^2}\right] = \frac{5}{4}\left[1 - \frac{2{,}4826}{(2{,}7237)^2}\right] = 0{,}8319$$

5. CONCLUSÃO

Apresentou-se neste trabalho as técnicas estatísticas de correlação que são fundamentais para a análise e construção de um instrumento de avaliação de qualidade.

6. REFERÊNCIAS

Chaves Neto, A., Turim; M. E., 2003, Análise de itens pela teoria clássica da avaliação e TRI em dados reais do ensino fundamental. In: Seminário Iasi de Estatística Aplicada, 9, Rio de Janeiro.

Cronbach, L. J., 1951, Coefficient alpha and the internal structure of testes. Psychometrika, v. 16, n. 3, pp. 297-333.

Ferguson, G. A.,1981, Statistical analysis in psychology and education, 5.ed., McGraw-Hill Book , New York, 549p.

Guilford, J. P., 1950, Fundamental statistics in psychology and education, 4.ed, McGraw-Hill Book, New York, 605p.

McNemar, Q., 1969, Psychological statistics, 4. ed, J. Wiley & Sons, New York, 529p.

7. DIREITOS AUTORAIS

CORRELATION ANALYSIS ADRESSED TO OBTAINING RELIABILITY AND DISCRIMINATION INDEX OF TEACHING ASSESSMENT TOOL

Sachiko Araki Lira
Instituto Paranaense de Desenvolvimento Econômico e Social – Ipardes
Rua Máximo João Kopp, 274 – Santa Cândida – Bloco 2 - CEP 82.630-900
Curitiba - PR
sachiko@onda.com.br

Anselmo Chaves Neto
Universidade Federal do Paraná
Programa de Pós-Graduação em Métodos Numéricos em Engenharia
Centro Politécnico – CEP 81.531-990
Curitiba – PR
anselmo@est.ufpr.br

Abstract. *Assessment is needed not just concerning production processes, but also when we want to know at what extent the clients are satisfied with the product they buy and the services they use. Regarding teaching, labor market increasingly requires competent professionals working efficiently. Thus, it is necessary to assess how students assimilate the contents taught and the quality of teachers' work, since teachers are indispensable. Chaves Neto and Turim (2003) compare the teaching process and assessment to the Shewhart cycle, also known as PDCA cycle (plan, do, check, act). That's to say; we have to plan and perform an action, assess its results and re-feed its planning continuously, seeking the continuous improvement of the teaching process. According to Chaves Neto and Turim, the basic conditions for an education measurement tool to be good are the following: validity, reliability (trustworthiness), objectivity and feasibility. On the other hand, the attributes of the test items are; difficulty extent and discrimination index. The Item Response Theory (IRT) includes models that relate student ability* (θ) *to item difficulty* (b) *and item discrimination index* (a). *Through this theory we can estimate the model parameters and classify both the items and the tests. The test reliability and the item discrimination extent can be estimated through the correlation coefficient. This work aims at showing different methods for reliability estimating: test-retest, parallel-form, split-half, Spearman-Brown, Kuder-Richardson and Cronbach Alfa.*

Keywords: *correlation, reliability.*