A TEORIA DA GENERALIZAÇÃO EM MEDIDAS EDUCACIONAIS EM PARALELO COM A TEORIA CLÁSSICA

Nicia Maria Besssa
Ph.D. em Educação, University of Pittsburgh,
M.A. em Psicologia, State University of Iowa.
Ex-professora da PUC/RJ e ex-pesquisadora do ISOP, Fundação Getúlio Vargas

Resumo
Este artigo apresenta uma introdução à teoria da generalização das medidas psicológicas e educacionais, sistematizada e extensamente detalhada por Cronbach e colaboradores em uma monografia publicada em 1972. Através de apresentação em paralelo com a versão mais conhecida da teoria clássica, procura-se mostrar a fundamentação comum e diferenças conceituais e metodológicas entre as duas. Esta introdução restringe seu foco a estudos dos erros de medida, os quais estão no cerne da teoria da generalização e são fundamentais na teoria clássica.
Palavras-chave: Medidas educacionais; teoria da generalização; teoria clássica.

Resumen
Este artículo presenta una introducción a la teoría de la generalización de las medidas psicológicas y educacionales, sistematizada y extensamente detallada por Cronbach y colaboradores en una monografía publicada en 1972. Son presentados en paralelo las versiones sintetizadas de la teoría de la generalización y de la teoría clásica como forma de destacar los puntos en común que tienen sus fundamentos, como también sus diferencias conceptuales y metodológicas. Esta introducción circunscribe su foco al estudio de los errores de las medidas, los cuales están en el centro de la teoría de la generalización y son fundamentales en la teoría clásica.
Palabras-clave: Medidas educacionales; teoría de la generalización; teoría clásica.

Abstract
This paper presents an introduction to generalizability theory, developed and extensively explained by Cronbach and other researchers in a monograph on the dependability of psychological and educational measurements published in 1972. Summary versions of generalizability theory and of the classic theory are presented in two parallel sections in order to show the common points in their fundamentals, as well as their conceptual and methodological differences. The focus is on studies of errors of measurement, which are in the heart of generalizability theory, and are fundamental to the classic theory.
Keywords: Educational measurement; generalizability theory; classic theory.

A concepção da teoria da generalização remonta à década de 60, quando Cronbach e colaboradores publicam os primeiros estudos nos quais introduzem conceitos e métodos de análise que focalizam o problema dos erros nas medidas psicológicas e educacionais de um modo que tem sido considerado como uma "liberalização" da teoria clássica (Cronbach et al.,1972; Feldt e Brennan,1993). No panorama teórico e metodológico da psicometria, a teoria da generalização se caracteriza por sua definição de erro de medida e pela ênfase na avaliação das fontes de erros de medida através da aplicação de processos da análise da variância (Sireci et al.,1998; Feldt e Brennan,1993).

Métodos de análise da variância foram introduzidos, há mais de cinqüenta anos, no estudo da fidedignidade das notas ou escores observados em medidas educacionais e psicológicas, no contexto do que se convencionou chamar de "teoria clássica" em psicometria. Entretanto, segundo Cronbach et al. (1972), os psicometristas demoraram a perceber as vantagens de investigar explicitamente múltiplos fatores que podem ser fontes de variabilidade nos resultados observados em medidas psicológicas ou educacionais, em parte por terem um enfoque que difere do enfoque da pesquisa experimental. Nesta última, de modo geral, se pessoas são objeto do experimento, elas são consideradas como fontes de "erro" pelo analista, interessado nos efeitos de tratamentos ou de fatores que especificam as condições em que a investigação se realiza; já o psicometrista está interessado principalmente em diferenças entre os sujeitos aos quais a medida se aplica e apenas secundariamente nas condições em que a observação se realiza.

Deve-se observar que a teoria clássica, além de ser utilizada por tantos anos na análise de resultados obtidos nas provas educacionais (veja-se por exemplo, o estudo das formas preliminares de uma prova de matemática por Bennet, Morley e Quardt,1998), tem servido também de ponto de partida para estudos que levaram à formulação de outras teorias, como a teoria da generalização e a teoria da resposta ao item. A análise de problemas que não são resolvidos pela teoria clássica levou à formulação de novos conceitos e teorias. Nesse sentido é interessante lembrar que Frederic Lord ao tempo que sistematizava, com Novick, as contribuições dos psicometristas à teoria clássica, também introduzia conceitos fundamentais, que se somaram aos estudos de Rasch, de Birnbaum e outros no

desenvolvimento da teoria que veio a ser conhecida como *Item Response Theory* (Hambleton,1993; Lord e Novick,1968).

Ao procurar contornar problemas que a teoria clássica suscita, a teoria da generalização introduz uma fundamentação conceitual nova e enfatiza a metodologia da análise da variância com o propósito de avaliar o efeito de diferentes fontes de erros de medida. Embora as diferenças conceituais sejam fundamentais entre as duas teorias, há uma base de idéias gerais comum a ambas – o que leva a se considerar que a teoria da generalização seja uma forma de "liberalização" da teoria clássica.

Nesta introdução à teoria da generalização procura-se deixar claro o sentido dessa "liberalização" através de uma apresentação paralela dos conceitos básicos assim como de métodos de análise da variância aplicados ao estudo dos erros de medida em uma e em outra teoria. Ao se restringir à teoria dos erros de medida, esta introdução não prejudica a compreensão da teoria, já que o estudo dos erros está no cerne da teoria da generalização e o estudo da fidedignidade das medidas psicológicas e educacionais é preocupação fundamental na teoria clássica.

Fidedignidade dos escores obtidos em provas educacionais – teoria clássica

Nesta seção apresenta-se resumidamente pontos de vista e conceitos fundamentais do que se tem chamado de "teoria clássica" das medidas psicológicas e educacionais, seguindo o desenvolvimento da teoria conforme é apresentada em textos de Gulliksen(1950) e de Lord e Novick (1968) ao tratarem das bases do conceito de fidedignidade.

- *Por que uma teoria estatística da medida em educação e psicologia?*

Do ponto de vista estritamento behaviorista, as respostas de um indivíduo às perguntas de uma prova de aritmética, por exemplo, podem ser consideradas como uma amostra da classe de respostas que ele daria, consistentemente, a estímulos semelhantes apresentados nas mesmas condições (Messick,1993). Assim, é possível argumentar que não é necessária uma teoria da medida para se prever o

comportamento desse indivíduo face a situações do mesmo teor, devendo-se empregar apenas um processo atuarial (Lord,1980).

Já a resposta daqueles que defendem a necessidade de uma teoria psicométrica reflete o ponto de vista dos que estabelecem relações entre o comportamento observado e constructos teóricos (Messick,1993; Lord,1980)

Sireci, Wainer e Braun (1998) argumentam que o valor de uma escala de medida não é determinado pelo processo experimental empregado, mas pelo modelo de medida – em particular pelo modelo utilizado ao se estimarem parâmetros, com base nos dados coletados experimentalmente.

E Lord e Novick(1968) explicam de modo didático :

> *"One reason we need to have a theory of mental testing is that mental test scores contain sizable errors of measurement. Another reason is that the abilities or traits that psychologists wish to study are usually not directly measurable ; rather, they must be studied indirectly, through measurements of other quantities".* (p.13)

É do comportamento observado, é da medida de propriedades das respostas dadas aos estímulos apresentados ao indivíduo (perguntas de uma prova de matemática, por exemplo) que se pode estudar, indiretamente, o conhecimento, a habilidade ou o traço psicológico que não se pode medir diretamente – ou seja, estudar o conceito teórico que se presume corresponder ao comportamento observado.

Quando se consideram os erros a que estão sujeitas as medidas educacionais e psicológicas, a teoria estatística tem a vantagem de permitir a interpretação e a explicação dos resultados obtidos empiricamente. Imagine-se, por exemplo, que um grupo de alunos obtenha escores baixos na escala de notas de uma prova de física e que em uma prova paralela, aplicada posteriormente, o grupo obtenha escores bem mais altos na mesma escala, dando a impressão de ter havido um ganho real, entre as duas medições, no conhecimento aferido. Entretanto, teoricamente é possível supor que os erros de medida tenham sido maiores na segunda aplicação e menores na primeira. Se a teoria permite deduzir essa possibilidade, o pesquisador pode-se precaver de antemão, planejando seu estudo de modo a poder eliminar interpretações indevidas.

➢ *Conceitos fundamentais da teoria clássica*

Há nuanças conceituais que permeiam conceitos básicos, tornando menos clara a concepção do que se denomina "teoria clássica" (Feldt e Brennan,1993; Lord e Novick, 1968). Entretanto, os estudos e as análises de resultados de provas educacionais têm, por muito tempo, partido de um conjunto de definições e de pressupostos que foram apresentados de modo sistemático, rigoroso e abrangente em textos como o Gulliksen (1950) e o de Lord e Novick (1968). O presente artigo deixa de lado tais diferenças conceituais e se concentra numa parte apenas desse conjunto: os fundamentos do estudo da fidedignidade dos escores das provas educacionais e psicológicas.

Em qualquer teoria das medidas psicológicas e educacionais reconhece-se que as fontes de erros de medida são diversas, desde variações inerentes ao desempenho do próprio indivíduo, até variações ambientais; além dessas, outras podem ser introduzidas, devidas a peculiaridades da amostra de estímulos (ou questões) apresentada em cada instrumento ou processo de medida e, em alguns casos, a peculiaridades do julgamento quando os resultados dependem da interpretação e do juízo de um examinador.

Lord e Novick (1968) formalizam a concepção de erro nas medidas educacionais ao definir o modelo linear em que o escore observado X_{gi} do indivíduo i na medida g é a soma do erro de medida E_{gi} e de um escore verdadeiro T_{gi}:

$$X_{gi} = T_{gi} + E_{gi}$$

Medidas repetidas, fixando-se o indivíduo i

O escore observado do indivíduo i na medida g na observação k (k = 1, 2, 3,...), X_{gki}, é concebido como uma variável aleatória que toma valores x_{gki} em uma seqüência de repetições de observações estatisticamente independentes do mesmo indivíduo i, nas mesmas condições de aplicação da medida g. A variável aleatória X_{gki} é definida no conjunto de todos os valores de x_{gki} que podem ser assim observados, considerando-se fixos o indivíduo i e a medida g.

O escore verdadeiro T_{gi} atribuído ao indivíduo i, na medida g, é definido como uma constante em uma distribuição teórica de valores observáveis X_{gki}, numa seqüência hipoteticamente infinita de repetições de medidas estatisticamente independentes do mesmo sujeito i, conservadas as mesmas condições de aplicação da medida g. Pressupõe-se que em cada repetição k (k=1, 2,...) da medida g do indivíduo i, o escore verdadeiro T_{gi} permaneça igual, ou seja, constante em toda a seqüência infinita de repetições.

Para cada repetição k (k = 1, 2, 3,...) da medida g do indivíduo i o escore observado X_{gki} é a soma do escore verdadeiro T_{gi} e do erro de medida E_{gki} :

$$X_{gki} = T_{gi} + E_{gki}$$

Nessas condições, define-se o escore verdadeiro T_{gi} como a expectância do escore observado X_{gki}, na seqüência infinita de repetições da medida g do indivíduo i, :

$$T_{gi} = E_k (X_{gki})$$

O erro de medida é definido pela diferença :

$$E_{gki} = X_{gki} - T_{gi}$$

sendo T_{gi} uma constante e sendo E_{gki} e X_{gki} variáveis aleatórias na seqüência de repetições da medida g do indivíduo i.

Considerando-se fixo o indivíduo i, teoricamente pode-se conceber medidas infinitamente repetidas de i, com um teste g, sendo o escore verdadeiro T_{gi} constante e sendo os erros de medida E_{gki} estatisticamente independentes; nessas condições, concebe-se uma distribuição teórica dos valores observados X_{gki} tal que :

$$F_{gki}(x_{gki}) = \text{Prob}(X_{gki} \leq x_{gki}) \qquad k = 1, 2, 3,...$$

Na distribuição $F_{gki}(x_{gki})$, a expectância dos escores observados X_{gki} é igual ao escore verdadeiro T_{gi} :

$$E_k (X_{gik}) = T_{gi}$$

Decorre do modelo de escore observado X_{gki} que a expectância dos erros E_{gki}, em $F_{gki}(x_{gki})$ é igual a zero :

$$E_k (E_{gik}) = 0$$

Fixado o indivíduo i e a medida g, a variância dos erros de medida em $F_{gi}(X_{gki})$ é igual à variância dos escores observados X_{gki} :

$$\sigma^2_{Egki} = E_k (X_{gki} - T_{gi})^2 = \sigma^2_{Xgki} \qquad k = 1, 2,,3,...$$

De modo geral, são definidas como medidas "repetidas" X_{gki} e $X_{gk'i}$ aquelas em que o escore verdadeiro do indivíduo i na medida g, na observação k, permanece igual na observação k'. Além disso, sendo k e k' medidas "repetidas", a estrutura probabilística do experimento não se altera de uma observação para outra e $F(E_{gki}) = F(E_{gk'i})$.

Variância de escores verdadeiros e observados em uma população P

Dado o interesse pelo estudo de diferenças individuais e pelo desempenho de grupos, a teoria das medidas educacionais e psicológicas focaliza a distribuição de escores observados em indivíduos selecionados aleatoriamente de uma população, aos quais se aplica uma medida g.

Na observação k, fixando-se a medida g, define-se uma variável aleatória X_{gki}, que toma valores observados x_{gki} para diferentes indivíduos selecionados aleatoriamente de uma população P, que se presume infinita.

O escore verdadeiro, T_{gi}, é constante na seqüência de repetições da medida g do indivíduo i. Entretanto, na população P, os escores verdadeiros podem variar de um indivíduo para outro. Sendo a medida g fixa, na observação k, na população P o escore verdadeiro, T_{gi} (i = 1, 2, 3,...), é definido como uma variável aleatória , que toma valores t_{gi} correspondentes aos escores verdadeiros de diferentes indivíduos .

Dado o modelo :

$$X_{gki} = T_{gi} + E_{gki} \qquad i = 1, 2, 3,...$$

o erro de medida é, também, uma variável aleatória em P, definida pela relação :

$$E_{gki} = X_{gki} - T_{gki} \qquad i = 1, 2, 3,...$$

Na distribuição de valores observados X_{gki} em uma população infinita de indivíduos, a expectância da variável aleatória X_{gki}, por todos os indivíduos da população P, é igual à média da distribuição dos escores verdadeiros :

$$E_P (X_{gki}) = E_P (T_{gi}) = \mu_{gX}$$

A expectância dos erros de medida, na população P é igual a zero :

$$E_P (E_{gki}) = E_P (X_{gki} - T_{gi}) = 0$$

Note-se, ainda, que a expectância dos erros de medida E_{gki} é igual a zero na subpopulação de P formada por indivíduos cujos escores verdadeiros T_{gi} (i = 1, 2, 3,...) são idênticos. Assim sendo, o coeficiente de regressão dos erros de medida E_{gki} em relação aos escores verdadeiros é igual a zero. E a correlação entre escores verdadeiros T_{gi} e erros de medida E_{gki} na população P é também igual a zero :

$$\rho_{(Egki, Tgi)} = 0$$

Numa seqüência de repetições da medida g , considerando-se fixo o indivíduo i, a variância dos erros de medida é a constante σ^2_{Egki}. Entretanto, essa variância pode tomar valores diferentes de um indivíduo para outro na população P. Define-se, então, uma variável aleatória σ^2_{Egki} que pode tomar valores diferentes quando são selecionados aleatoriamente indivíduos da população P.

Na população P, a variância dos erros de medida, σ^2_{Eg}, é igual à expectância, em P, da variância da distribuição teórica de erros, em todas as repetições da medida g do indivíduo i :

$$\sigma^2_{Eg} = E_P (\sigma^2_{Egki})$$

Ou seja :

$$E_P E_k (E^2_{gki}) = E_P E_k (X_{gki} - T_{gi})^2 = E_k E_P (X_{gki} - T_{gi})^2 = E_k (\sigma^2_{Eg}) = \sigma^2_{Eg}$$

sendo: k = 1, 2, 3,...
i = 1, 2, 3,...

Como decorrência das definições e das relações acima tem-se que, na população infinita de indivíduos submetidos à medida g, a variância dos escores observados σ^2_{Xg} é igual à soma da variância dos escores verdadeiros σ^2_{Tg} e da variância dos erros de medida σ^2_{Eg}:

$$\sigma^2_{Xg} = \sigma^2_{Tg} + \sigma^2_{Eg}$$

O conceito de fidedignidade

A fidedignidade dos escores X_{gki} dos indivíduos de uma população P, em uma prova g, é definida como a proporção da variância dos escores verdadeiros na variância dos escores observados. Quanto maior a proporção da variância dos escores verdadeiros na variância dos escores observados, maior a fidedignidade das medidas observadas na aplicação de g na população P.

Sendo igual a zero a correlação entre erros de medida e escores verdadeiros, $\rho_{(Egik,\ Tgi)}$, a expectância, na população P, do produto dos erros de medida e dos escores verdadeiros é também igual a zero:

$$E_P\ (T_{gi}.\ E_{gki}) = 0$$

E a covariância entre os escores observados e os escores verdadeiros, $COV_{Xg,\ Tg}$, é igual à variância dos escores verdadeiros σ^2_{Tg}:

$$COV_{Xg,\ Tg} = E_P\ (X_{gki}.T_{gi}) - E_P\ (X_{gki})\ E_P\ (T_{gi}) = \sigma^2_{Tg} + E_P\ (E_{gki}.\ T_{gi}) = \sigma^2_{Tg}$$

A correlação entre escores observados e verdadeiros é, então, definida:

$$\rho_{(Xgki,\ Tgi)} = COV_{Xg,\ Tg} / \sigma_{Xg}.\sigma_{Tg} = \sigma^2_{Tg} / \sigma_{Xg}\ .\ \sigma_{Tg}$$

O quadrado da correlação entre escores verdadeiros T_{gi} e escores observados X_{gki} é igual à proporção da variância dos escores verdadeiros na variância dos escores observados:

$$\rho^2_{Xgki,\ Tgi} = \sigma^2_{Tg}\ .\ \sigma^2_{Tg} / \sigma^2_{Xg}\ .\ \sigma^2_{Tg} = \sigma^2_{T\,g} / \sigma^2_{Xg}$$

Ou seja, a fidedignidade dos escores observados, na população P, na prova g, é definida pelo quadrado da correlação entre escores verdadeiros e observados.

A imprecisão dos escores X_{gki}, observados em g, na população P, é também expressa pela variância dos erros de medida σ^2_{Eg}. Ou, ainda, pelo desvio-padrão dos erros de medida, σ_{Eg} – este último denominado erro padrão da medida. A interpretação do erro-padrão da medida requer atenção à escala de medida empregada, e requer a introdução de pressupostos outros além daqueles introduzidos para o desenvolvimento de grande parte da teoria.

➢ *O conceito de medidas paralelas*

Interessa fixar o conceito de medidas paralelas na teoria clássica para melhor situar diferenças entre esta e a teoria da generalização.

Na apresentação de Gulliksen(1950) e na de Lord e Novick (1968) não se considera necessário, para o desenvolvimento da maior parte da teoria clássica, o pressuposto de idênticas distribuições de erros de medida que integra a conceituação de medidas "repetidas". Introduz-se o conceito de medidas "distintas", g e h, tais que :

$$E_k (X_{gki} \mid x_{hki}) = E_k (X_{gki}) \quad \text{e} \quad E_k (X_{hki} \mid x_{gki}) = E_k (X_{hki})$$

Esta definição de medidas "distintas" serve de base ao conceito de medidas "paralelas", importante na teoria da fidedignidade dos escores obtidos em provas psicológicas e educacionais já que, em muitos casos, essas medidas não podem ser aplicadas ao mesmo indivíduo repetidamente e, em outros, a aplicação repetida pode refletir-se em alterações no desempenho do próprio indivíduo.

Tomando- se duas medidas "distintas", g e h, aplicadas a uma população infinita de indivíduos, tem-se duas distribuições de erros de medida; a correlação entre essas duas variáveis aleatórias é igual a zero, assim como também é igual a zero a correlação entre erros na medida g e escores verdadeiros na medida h :

$$\rho_{(Egi,\ Ehi)} = 0 \quad \text{e} \quad \rho_{(Egi,\ Thi)} = 0$$

Conforme Lord e Novick (1968), medidas paralelas X_{gi} e X_{hi} são definidas como medidas "distintas", sendo iguais os respectivos escores verdadeiros $T_{gi} = T_{hi}$ e sendo iguais as variâncias dos erros de medida σ^2_{Eg} e σ^2_{Eh} na população de indivíduos P.

As medidas "distintas" X_{gi} e X_{hi} são paralelas se, para cada subpopulação de P:

$$T_{gi} = T_{hi} \qquad e \qquad \sigma^2_{Eg} = \sigma^2_{Eh}$$

Como decorrência :

$$E_P(X_{gi}) = E_P(X_{hi}) \qquad e \qquad \sigma^2_{Xg} = \sigma^2_{Xh}$$

Sendo paralelas as medidas X_{gi}, X_{hi} e X_{zi}, são iguais as intercorrelações e as correlações com outra medida qualquer X_{yi} :

$$\rho(X_{gi}, X_{hi}) = \rho(X_{gi}, X_{zi}) = \rho(X_{hi}, X_{zi})$$

$$\rho(X_{gi}, X_{yi}) = \rho(X_{hi}, X_{yi}) = \rho(X_{zi}, X_{yi})$$

Não obstante as dificuldades práticas envolvidas na construção de formas paralelas de testes educacionais, o conceito de medidas paralelas ocupa lugar central na pesquisa psicométrica no contexto da teoria clássica. Com a introdução do conceito, pode-se definir a fidedignidade dos escores, numa população P, pela correlação entre as medidas paralelas X_{gi} e X_{hi}. Substituindo X_{gi} e X_{hi} por $(T_{gi} + E_{gi})$ e $(T_{hi} + E_{hi})$ respectivamente, tem-se que a covariância dos escores observados em medidas paralelas é igual à variância dos escores verdadeiros:

$$\text{cov}\, X_g X_h = E_P\{[(T_{gi} + E_{gi}) - E_P(T_{gi} + E_{gi})][(T_{hi} + E_{hi}) - E_P(T_{hi} + E_{hi})]\} =$$

$$= \sigma^2_{Tg} = \sigma^2_{Th}$$

Sendo iguais as variâncias, a correlação entre as medidas paralelas g e h expressa a proporção da variância verdadeira na variância observada total, ou seja, é igual ao coeficiente de fidedignidade:

$$\rho(X_{gi}, X_{hi}) = \rho^2(X_{gi}, T_{gi}) = \rho^2(X_{hi}, T_{hi}) = \sigma^2_{Tg} / \sigma^2_{Xg} = \sigma^2_{Th} / \sigma^2_{Xh}$$

Embora não seja essencial para o desenvolvimento da teoria, o conceito de medidas paralelas é fundamental para a pesquisa psicométrica.

➢ *Aplicações da análise da variância*

Os psicometristas, ao empregarem a metodologia da análise da variância para estudar a fidedignidade dos escores observados na aplicação de medidas educacionais, além de obter estimativas do coeficiente de fidedignidade visam a analisar efeitos de fontes diferentes sobre a variância dos escores observados.

No contexto da teoria clássica pode-se empregar o conceito de medidas "repetidas" de cada indivíduo i selecionado aleatoriamente de uma população P, com a medida g, para estimar a variância verdadeira e a variância devida a erros de medida.

Em um exemplo simples, pode-se propor um planejamento experimental para estimar a variância dos escores verdadeiros σ^2_{Tg} e a variância dos erros de medida σ^2_{Eg}. O experimento consiste em se aplicar a medida g a uma amostra aleatória de indivíduos de uma população P, sendo que para cada indivíduo i haverá uma seqüência de "repetições" da medida g. Supõe-se que o processo experimental resguarde a independência dos erros de medida E_{gik} na seqüência de observações da medida g do indivíduo i.

No modelo :

$$X_{gki} = \mu + (T_{gi} - \mu) + E_{gki} \qquad k = 1, 2, 3,...$$

$$i = 1, 2, 3,...$$

X_{gki}, T_{gi} e E_{gki} são variáveis aleatórias e a média da população P, representada por μ, é a expectância dos escores observados:

$$E_p E_k (X_{gki}) = E_p (T_{gi}) = \mu$$

Na análise da variância obtém-se a estimativa da variância intra-indivíduos, σ^2_{Eg}, devida a erros de medida já que se refere ao conjunto das distribuições de erros nas "repetições" da medida g para cada um dos indivíduos – sendo o escore verdadeiro do indivíduo i constante em todas as "repetições", a variabilidade dos escores observados em cada uma dessas distribuições deve-se aos erros de medida. Obtém-se também a estimativa da variância entre indivíduos, formada pela variância dos erros de medida σ^2_{Eg} e pela variância verdadeira σ^2_{Tg}.

Para a obtenção dessas estimativas pela análise da variância, são introduzidos os pressupostos usuais pertinentes ao modelo. Quando os objetivos da pesquisa não incluem testes de significância, não são considerados necessários os pressupostos relativos a distribuições normais de diferenças devidas a erros, ou de outros efeitos focalizados no modelo.

Modelos lineares mais ou menos complexos têm sido empregados na aplicação da análise da variância aos resultados de experimentos em que medidas paralelas são aplicadas a amostras de indivíduos de uma população. Conforme seja o arranjo experimental, pode-se estimar as variâncias de efeitos de diferentes fontes sobre a variabilidade dos escores obtidos.

Exemplos didáticos desse tipo de análise são encontrados em textos como os de Lord e Novick (1968), de Winer (1971), ou de Stanley (1971). Já Ebel (1967) se serve de casos de notas atribuídas por juizes a trabalhos de alunos para analisar detalhadamente a importância do emprego da análise da variância para se estimar a contribuição de fontes diversas que podem inflacionar a variância devida a erros de medida – neste caso, as diferenças entre juizes.

No Brasil, os estudos de Vianna (1976) e de Bessa (1986) focalizam a fidedignidade de notas atribuídas a redações. O primeiro avalia as diferenças entre médias de notas atribuídas por vários avaliadores. Ao avaliar a fidedignidade das notas obtidas em redações por um grupo de alunos, o segundo utiliza um modelo de componentes da variância para pesquisar diferenças entre os julgamentos dos professores como fonte de variabilidade entre os escores observados. Conforme se defina o modelo, a variância entre avaliadores poderá ser considerada como parte da variância devida a erros, ou pode ser subtraída da variância intra-indivíduos.

Em suma, ao empregar a análise da variância nos estudos de fidedignidade de escores obtidos em medidas educacionais e psicológicas, no contexto da teoria clássica, os psicometristas usam modelos adequados a planejamentos experimentais em que os sujeitos são submetidos a medidas "repetidas", ou a aplicações de medidas "paralelas", dentro de um esquema conceitual em que o erro é definido como a diferença entre o escore observado e uma constante, que é o escore verdadeiro – este definido como a expectância dos escores observados em repetições da mesma medida, nas mesmas condições, a um indivíduo.

Ao avaliar a fidedignidade dos escores observados empregando-se a análise da variância de medidas repetidas, o coeficiente obtido pode refletir uma estimativa da variância dos erros de medida na qual se integra a variância devida a diferenças entre médias de fonte presente no experimento, mas não explicitada no modelo. Ou, pelo contrário, quando no modelo é explicitado o efeito de certa fonte de variabilidade, introduzida no experimento, somando-se aos efeitos das repetições da mesma medida sobre cada indivíduo, pode-se subtrair a variância entre essas médias da variância intra-sujeitos; neste caso, a estimativa do coeficiente de fidedignidade reflete apenas a variabilidade devida a efeitos residuais.

Por exemplo, em um experimento em que a cada redação de um grupo de examinandos sejam atribuídas notas por quatro avaliadores, pode-se representar por X_{gi} a nota do indivíduo i atribuída pelo avaliador g , em um modelo em que π_i é o escore verdadeiro de i e η_{gi} simboliza o erro de medida :

$$X_{gi} = \pi_i + \eta_{gi}$$

Neste caso, a variância intra-sujeitos engloba a variância entre avaliadores e variância devida a outra fontes de erros de medida em σ^2_{η}.

De outro lado, pode-se explicitar em um modelo os efeitos α_g correspondentes às diferenças entre as médias das notas atribuídas pelos avaliadores:

$$X_{gi} = \pi_i + \alpha_g + \eta_{gi}$$

Ao estimar a fidedignidade das notas observadas, pode-se subtrair da variância intra-sujeitos a variância devida a diferenças entre as médias dos avaliadores; neste caso, σ^2_{η} reflete apenas diferenças entre repetições das medidas (veja-se. por exemplo, Winer, 1971, p. 284-289).

A confiabilidade dos escores – teoria da generalização

Uma apresentação, ainda que sumária, da teoria da generalização requer não só a definição de conceitos fundamentais como também um entendimento da maneira pela qual a teoria recorre

à metodologia da análise da variância para estudar as fontes de erros introduzidos nas medidas educacionais e psicológicas. Para tanto, a comparação com os conceitos e os métodos de análise empregados no contexto da teoria clássica pode ser esclarecedora.

➢ *Conceitos fundamentais*

Na teoria da generalização o termo ***população*** se aplica ao conjunto de indivíduos que são objeto de mensuração. Já o termo ***universo*** é reservado para designar um conjunto de condições em que as medidas são realizadas – condições essas que são semelhantes entre si, ainda que não sejam inteiramente iguais. Por exemplo, são condições de medida : formas diversas de uma prova de Matemática que sejam semelhantes, mas que não podem ser consideradas formas paralelas no sentido compreendido na teoria clássica – pode acontecer que essas provas tenham níveis de dificuldade desigual.

Condições de medida semelhantes constituem uma ***faceta*** ("facet"), na terminologia da teoria da generalização. Numa pesquisa, várias ***facetas*** podem ser introduzidas simultaneamente: podem ser juizes que atribuem notas a provas, podem ser diferentes estilos de gestão escolar, podem ser regiões de um país, podem ser diferentes tipos de prova escolar etc. As diferentes ***facetas*** introduzidas numa pesquisa integram o ***universo de observações admissíveis.*** Em uma pesquisa em que quatro professores atribuem notas a cada uma das formas de um teste de Física aplicadas a um grupo de alunos, há duas ***facetas*** que integram o ***universo de observações admissíveis***: professores e formas do teste de Física.

Quando interpreta o resultado apresentado por um indivíduo em um teste, o psicometrista tem em vista que aquele escore, obtido em uma prova determinada, aplicada em uma situação particular, em certa ocasião, é apenas uma amostra dos resultados que o mesmo indivíduo pode apresentar e que faz parte de um conjunto de provas, situações e ocasiões semelhantes - ou seja, faz parte de um ***universo de observações admissíveis.***

Considere-se, por exemplo, que se aplique ao indivíduo i três formas de uma prova de compreensão de leitura, cada uma com um texto sobre um tema diferente; formando os temas de leitura uma ***faceta,*** que se presume ter infinitas condições dentro do ***universo de observações admissíveis***; concebe-se que os temas sejam uma amostra

aleatória do universo de temas de leitura. Considera-se também, que o escore observado, X_{it}, em uma forma da prova, seja uma amostra dos escores do indivíduo i nesse universo. Interessa investigar o erro de medida ao se interpretar o escore observado X_{it} como representativo do ***escore universal*** de i, definido como a expectância do escore observado, por todas as condições da ***faceta*** t:

$$\mu_i = E_t (X_{ti})$$

A concepção de erro de medida

Na teoria da generalização a definição do erro de medida depende da maneira como é considerado o resultado observado. No exemplo acima, o escore X_{it} pode ser tomado pelo psicometrista como uma estimativa do escore universal do indivíduo i (ou seja, de μ_i), sem qualquer referência à distribuição de escores na população de indivíduos na qual i está integrado. Neste caso, define-se o erro de medida, cujo símbolo é Δ, na teoria da generalização:

$$\Delta_{ti} = X_{ti} - E_t(X_{ti}) = X_{it} - \mu_i$$

Altera-se o sistema de notação quando o escore observado do indivíduo i é definido como uma média aritmética (ou como uma soma) dos escores obtidos em várias condições de uma ***faceta*** – por exemplo, uma média aritmética dos pontos obtidos em quatro temas de leitura apresentados em uma prova. Em casos como este, na notação da teoria da generalização, uma letra maiúscula representa a ***faceta*** estudada. No exemplo, a letra T substituiria a letra minúscula t em toda a notação empregada na investigação: X_{Ti} representaria a média aritmética das notas do indivíduo i nos quatro temas de leitura, sendo o ***escore universal***, μ_i, definido pela expectância dessa nota média, por todos os temas de leitura:

$$\mu_i = E_T (X_{Ti})$$

Há casos em que o psicometrista está interessado na posição do escore obtido por um indivíduo i em relação à distribuição de escores na ***população*** de indivíduos da qual foi selecionado – por exemplo, em relação à média dos escores dessa ***população***. Seja, por exemplo, X_{fi} a nota de um aluno i, selecionado da população P, na forma f

de um teste de Aritmética ; para o pesquisador interessa expressar a posição de i na distribuição de notas na ***população*** P, comparando essa nota com a média das notas na ***população*** P – ou seja, o interesse do pesquisador está em $X_{fi} - E_i(X_{fi})$. Para avaliar o erro de medida compara-se a posição do escore observado ($X_{fi} - E_i(X_{fi})$) com a posição do **escore *universal*** do indivíduo i, μ_i em relação à média na **população** e no **universo de observações admissíveis**, μ – isto é, com $\mu_i - \mu$. Ou seja, sendo:

$$\mu_i = E_f(X_{fi})$$

$$\mu_f = E_i(X_{fi})$$

$$\mu = E_i E_f(X_{fi})$$

define-se o erro de medida, cujo símbolo é δ, na teoria da generalização :

$$\delta_{fi} = [X_{fi} - E_i(X_{fi})] - [E_f(X_{fi}) - E_i E_f(X_{fi})] = (X_{fi} - \mu_f) - (\mu_i - \mu)$$

O coeficiente de generalização

A fidedignidade dos escores de uma medida psicológica ou educacional, na teoria da generalização, é concebida de modo análogo àquele em que se define na teoria clássica: é a proporção da variância dos ***escores universais*** dos indivíduos de uma ***população*** (em correspondência à variância verdadeira da teoria clássica) na expectância da variância total observada.

O coeficiente de generalização, representado por $E\rho^2$, é definido pela relação entre a variância dos ***escores universais***, σ^2_i, e a variância total, que é a soma da variância devida aos erros de medida, σ^2_δ, e da variância dos escores universais σ^2_i :

$$E\rho^2 = \sigma^2_i / \sigma^2_i + \sigma^2_\delta$$

➢ *Estudos G e D*

De modo geral, estudos preliminares são realizados durante a construção de um instrumento ou de um procedimento de medida educacional ou psicológica de modo a se avaliar as características psicométricas e demais requisitos para satisfazer os critérios determinados pelo investigador. Na teoria da generalização essas pesquisas são formalizadas, de maneira particular, ao se caracterizar a diferença entre os chamados ***estudos G*** (estudos de generalização) e ***estudos D*** (estudos de decisão) (Cronbach et al., 1972). Os ***estudos G*** são preliminares e são parte do desenvolvimento do instrumento ou do procedimento de medida. Os ***estudos D*** se apoiam nos resultados dos ***estudos G*** no sentido de usar seja resultados obtidos, seja instrumentos ou procedimentos de medida desenvolvidos, seja de usar planejamentos de pesquisa de maneira a melhor servir aos objetivos do psicometrista.

Os ***estudos G*** têm por finalidade desenvolver instrumentos ou procedimentos de medida que sirvam especificamente aos objetivos de um ou mais ***estudos D*** ; já estes últimos têm por objetivo estudar indivíduos ou grupos de indivíduos, ou estudar relações entre variáveis que caracterizam grupos de indivíduos.

Em um ***estudo D*** o psicometrista pretende generalizar o escore observado do indivíduo i por todo um universo de condições em que a observação se opera. Denomina-se ***universo de generalização*** esse universo de condições em que o psicometrista está interessado ao generalizar a observação de um indivíduo ou de um grupo de indivíduos. Assim sendo, o ***estudo G*** deve incluir no ***universo de observações admissíveis*** aquelas ***facetas*** que fazem parte do ***universo de generalização***. Mais ainda, o planejamento experimental do ***estudo G*** deve permitir que sejam estimados componentes da variância tais que o estudo possa ser útil a um ou mais ***estudos D*** particulares.

Numa série de estudos realizados por Shavelson, Baxer e Gao (1993) encontram-se exemplos de ***estudos G*** que avaliam a confiabilidade de provas de desempenho de alunos em tarefas de laboratório de Ciências e em questões abertas de Matemática ;tanto os trabalhos, como as respostas livres dos alunos na prova de Matemática recebem notas atribuídas por diversos professores em mais de uma ocasião. Em relação a cada uma dessas matérias escolares, ***estudos G*** são apresentados usando diferentes modelos de análise que incluem

duas ou três das seguintes ***facetas***: professores-avaliadores, tarefas e ocasiões. Com base em estimativas de componentes da variância obtidas em cada ***estudo G,*** são estimadas as variâncias devidas a erros de medida (σ^2_δ e σ^2_Δ) quando se altera o número de tarefas ou de professores-avaliadores – ou seja, os modelos usados nos ***estudos G*** permitem fazer simulações para que o psicometrista tome decisões em relação ao futuro emprego de tais provas em situações práticas. Por exemplo, com um modelo em que entram apenas duas ***facetas*** – tarefas e professores – como fatores cruzados com indivíduos, os escores em uma das provas de Ciências, constituída por cinco (5) trabalhos de laboratório, apresentam como maior fonte de erro de medida a interação entre indivíduos e tarefas. Num ***estudo D***, usando os valores obtidos nesse ***estudo G,*** estima-se que deveriam ser usadas oito (8) tarefas de laboratório de Ciências para diminuir convenientemente o impacto dos erros de medida, no caso de se empregar σ^2_δ; estima-se também que deveriam ser dez (10) as tarefas de laboratório, no caso de se usar σ^2_Δ. Nesse ***estudo D*** em particular, o psicometrista deixa de lado a ***faceta*** denominada ocasiões, pois está interessado em um ***universo de generalização*** que abrange apenas as condições que integram duas das ***facetas***: tarefas e professores ; assim sendo, o **estudo G** correspondente inclui somente essas duas ***facetas*** no ***universo de observações admissíveis***, já que o objetivo do psicometrista, no **estudo D**, é verificar o erro de medida cometido ao considerar a nota X_{tpi} atribuída pelo professor p, em uma tarefa t, ao indivíduo i, como representativa do escore de i em todas as tarefas e segundo o julgamento de todos os professores que integram o ***universo de generalização*** – ou seja, representativa de μ_i.

➢ *Aplicações da análise da variância*

Diversos modelos lineares são empregados em análises da variância, sejam univariadas ou multivariadas, em **estudos G** e ***D***. Em geral, trata-se de modelos de componentes da variância, ou de modelos mistos ; ocasionalmente são usados modelos hierárquicos. A análise da variância aplicada aos dados obtidos nos estudos empíricos permite ao psicometrista avaliar os efeitos das fontes de erros de medida em cada caso específico. Cronbach et al. (1972) descrevem detalhadamente uma quantidade de modelos, de planejamentos

experimentais e de procedimentos de análise para estimar componentes da variância; Feldt e Brennan (1993) apresentam um resumo bastante completo dos tipos de modelos mais utilizados e de estimadores de componentes da variância. Alguns estudos de Brennan e colaboradores (1995) e de Bennet et al.(1993) servem para ilustrar aplicações da teoria.

Brennan et al. (1995) apresentam análises univariadas e multivariadas em uma série de estudos em que avaliam características psicométricas de formas experimentais de dois testes de linguagem. No experimento, cada indivíduo ouve doze (12) mensagens gravadas e anota o que ouve durante a apresentação de cada uma (teste L) ;após cada apresentação o examinando produz um resumo escrito da mensagem, com base nas próprias anotações (teste W). Três avaliadores atribuem notas de zero (0) a cinco (5) às anotações relativas a cada mensagem (teste L) ; outros três avaliadores fazem o mesmo com cada resumo escrito (teste W) produzido por cada um dos 50 indivíduos. O experimento consiste, pois, em submeter ao teste L e ao teste W um grupo de 50 pessoas, que recebem um total de 36 notas por teste; nesse arranjo, as mensagens e os avaliadores são ***facetas,*** cujos efeitos sobre os erros de medida importa estudar. No ***estudo G***, pressupõe-se que as ***facetas*** de mensagens e de avaliadores sejam infinitamente grandes no **universo de *observações admissíveis***, assim como a **população** de indivíduos da qual os examinandos são considerados uma amostra aleatória. Em uma análise de componentes da variância, em um dos ***estudos G***, relativo a uma das formas experimentais do teste L, a nota atribuída ao indivíduo i, na mensagem t, pelo avaliador r é representada pelo modelo :

$$X_{tri} - \mu + (\mu_i - \mu) + (\mu_t - \mu) + (\mu_r - \mu) + (\mu_{ti} - \mu) + (\mu_{ri} - \mu) + (\mu_{tr} - \mu) + (\mu_{tri} - \mu)$$

em que

$$\mu = E_t E_t E_r (X_{tri}) \qquad \text{e} \qquad \mu_i = E_t E_r (X_{tri})$$

Os demais termos são efeitos dos fatores "mensagens" (ou "tarefas") e "avaliadores" e das interações, sendo que o efeito residual ($\mu_{tri} - \mu$) envolve a interação de pessoas, mensagens, avaliadores e fontes de erros de medida não identificadas no modelo. Na análise univariada, a variância total das notas observadas (X_{tri}) no teste L é integrada pelos componente :

$$\sigma^2_{Xtri} = \sigma^2_i + \sigma^2_t + \sigma^2_r + \sigma^2_{ti} + \sigma^2_{ri} + \sigma^2_{tr} + \sigma^2_{tri}$$

Nesse ***estudo G*** são estimados os componentes da variância dos escores observados. Os resultados apresentados indicam diferenças consideráveis entre os indivíduos, conforme a estimativa de σ^2_i. Note-se que, na teoria da generalização, σ^2_i representa variância dos escores universais (μ_i) e que os demais componentes da variância são incluídos na variância devida a erros de medida. No caso particular deste estudo, a variância devida à interação entre pessoas e mensagens (ou tarefas), representada por σ^2_{ti}, contribui fortemente para a variância dos erros de medida, indicando que as notas dos indivíduos diferem bastante em relação à dificuldade de cada tarefa no teste L.

Com base nas estimativas dos componentes da variância, são apresentados ***estudos D*** nos quais a definição da nota atribuída a cada examinando é diferente daquela empregada nos ***estudo G***. Nos ***estudos D***, a nota atribuída ao indivíduo i é concebida como uma média das notas atribuídas a n'_t mensagens por n'_r avaliadores. Nos ***estudos D*** são feitas simulações em que o número de mensagens n'_t ou o número de avaliadores n'_r varia, ao se calcular as notas médias de cada indivíduo. Conforme a notação da teoria da generalização, T e R representam as ***facetas*** de mensagens e de avaliadores respectivamente no modelo:

$$X_{TRi} = \mu + (\mu_i - \mu) + (\mu_T - \mu) + (\mu_R - \mu) + (\mu_{Ti} - \mu) + (\mu_{Ri} - \mu) + (\mu_{TR} - \mu) + (\mu_{TRi} - \mu)$$

A variância total dos escores médios (X_{TRi}) atribuídos aos indivíduos, ou seja dos escores observados nesse ***estudo D*** é representada por:

$$\sigma^2_{XTRi} = \sigma^2_i + \sigma^2_T + \sigma^2_R + \sigma^2_{Ti} + \sigma^2_{Ri} + \sigma^2_{TR} + \sigma^2_{TRi}$$

A relação entre os componentes da variância referentes ao modelo do ***estudo G*** e os referentes ao modelo do ***estudo D*** é definida por Brennan et al. (1995):

$$\sigma^2_T = \sigma^2_t / n'_t \qquad \sigma^2_R = \sigma^2_r / n'_r$$
$$\sigma^2_{Ti} = \sigma^2_{ti} / n'_t \qquad \sigma^2_{Ri} = \sigma^2_{ri} / n'_r$$
$$\sigma^2_{TR} = \sigma^2_{tr} / n'_t n'_r \qquad \sigma^2_{TRi} = \sigma^2_{tri} / n'_t n'_r$$

Nesse ***estudo D*** a variância σ^2_i é a mesma do ***estudo G*** – é a variância dos ***escores universais***. O ***universo de generalização*** é concebido como sendo formado por todas as medidas - aplicadas a amostras aleatórias de indivíduos da população do ***estudo G*** - em que amostras aleatórias de n'_t mensagens e de n'_r avaliadores são selecionadas dos respectivos universos e em que os escores individuais são notas médias, como definidas no modelo.

Se o interesse do pesquisador é avaliar diferenças entre indivíduos, não fazem parte da variância dos erros de medida, σ^2_δ, aquelas condições que afetam igualmente a todos. Note-se que se trata da variância dos erros tipo δ que, neste caso, é definida :

$$\sigma^2_\delta = \sigma^2_{Ti} + \sigma^2_{Ri} + \sigma^2_{TRi}$$

Se o foco da pesquisa é o valor do escore obtido pelo examinando sem referência à média do grupo, a variância dos erros tipo Δ é definida:

$$\sigma^2_\Delta = \sigma^2_T + \sigma^2_R + \sigma^2_{Ti} + \sigma^2_{Ri} + \sigma^2_{TR} + \sigma^2_{TRi}$$

Para diferentes valores de n'_t e de n'_r, o estudo apresenta comparações entre estimativas de variâncias dos erros de medida σ^2_δ e σ^2_Δ. No teste L, quanto menor o número de avaliadores, maior o número de tarefas deve ser usado para diminuir a variância dos erros de medida σ^2_δ e σ^2_Δ. Comparação análoga é feita para estimar o número de tarefas (n_t) e de avaliadores (n_r) para se obter mais altos ***coeficientes de generalização*** :

$$E\rho^2 = \sigma^2_i / \sigma^2_i + \sigma^2_\delta$$

e ***índices de confiabilidade***, cuja representação no trabalho de Brennan et al. é:

$$\phi = \sigma^2_i / \sigma^2_i + \sigma^2_\Delta$$

Como esperado, os índices de generalização crescem quando o número de avaliadores passa de um (1) para três (3) e o número de tarefas passa de seis (6) para doze (12).

Os resultados das análises multivariadas dos ***estudos G*** e ***D*** – que se aplicam aos testes L e W – são semelhantes. Nos ***estudos G***, como os avaliadores formam grupos diferentes para os testes L e W,

são nulas as covariâncias que envolvem a ***faceta*** "avaliadores". As demais covariâncias estimadas permitem que, no correspondente ***estudo D*** sejam avaliados os efeitos de alterações no número de tarefas ou no número de avaliadores. Tal como na análise univariada, no ***estudo D*** supõe-se que a cada examinando i corresponda uma nota, X_{TRi}, que é a média dos escores em n'_t tarefas, atribuídas por n'_r avaliadores. Investiga-se, então, a confiabilidade da diferença entre duas notas (X_{TRi} no teste L menos X_{TRi} no teste W) de cada examinando. Em relação a essas diferenças, observa-se que a estimativa da variância dos erros Δ é tanto menor conforme n'_t varia de seis (6) a doze (12) e n'_r varia de um (1) a três (3). Além disso, nesse ***estudo D***, a análise multivariada focaliza questões relativas à correlação entre erros do tipo Δ, ao fazer variar as condições de amostragem das tarefas.

Os ***estudos G*** e ***D*** conduzidos por Bennett e Rock (1993) têm por objetivo avaliar características psicométricas de um teste de formulação de hipóteses (teste F-H) apresentado a cada examinando na tela de microcomputador; cada questão consiste em um problema para cuja solução o examinando deve formular uma hipótese. O teste F-H contém oito (8) questões, sendo que quatro das quais devem ser respondidas, no máximo, com sete (7) palavras e outras quatro questões, com quinze (15) palavras. O experimento envolve 30 indivíduos, cujas respostas a cada questão recebem notas entre zero (0) e quinze (15), atribuídas independentemente por cinco (5) avaliadores. Foram estimados os componentes da variância dos escores Y_{jki}, definidos no modelo :

$$Y_{jki} = \mu + R_j + Q_k + \pi_i + RQ_{kj} + \pi R_{ji} + \pi Q_{ki} + \pi RQ_{jki}$$

em que Y_{jki} é a nota do indivíduo i atribuída pelo avaliador j na questão k. A média geral por todas as ***facetas*** e por toda a ***população*** é representada por μ. Pressupõe-se que os efeitos das ***facetas*** avaliador (R) e questão (Q) sejam amostras aleatórias de infinitas diferenças no ***universo de observações admissíveis***, assim como os efeitos dos indivíduos (π) sejam uma amostra aleatória de diferenças entre ***escores universais*** de uma ***população*** infinita.

No ***estudo G*** são apresentados os resultados de duas análises univariadas separadamente: uma para a situação de limite de sete (7)

palavras e outra para a de limite de quinze (15) palavras. Os dois ***estudos G*** mostram fortes diferenças entre indivíduos; apenas o efeito da interação entre pessoas e questões (πQ_{ki}) aparece com uma contribuição importante para a variância dos erros de medida (neste caso, em particular, σ^2_{Δ}).

Os ***estudos D*** indicam que, para atingir um ***coeficiente de generalização*** de 0,80 o teste F-H deveria conter duas (2) ou três (3) questões cujas respostas fossem limitadas a sete (7) palavras ; para a situação de limite de quinze (15) palavras, deveria conter entre três(3) e quatro (4) questões.

Com esses dois exemplos, pode-se ter uma visão, ainda que superficial, da maneira pela qual a teoria da generalização usa a modelagem e os processos da análise da variância para estudar as fontes de erros de medida e como, nos ***estudos D***, ela explora alterações em modelos e em condições experimentais com o fim de avaliar formas de aperfeiçoar instrumentos e processos de medida em psicologia e educação.

Considerações finais

Na apresentação paralela da teoria clássica e da teoria da generalização pode-se perceber claramente pontos comuns às duas e diferenças que se refletem no estudo da confiabilidade dos escores observados nas medidas psicológicas e educacionais.

Em ambas, a concepção de um modelo teórico de um escore ou nota observada em uma medida do um indivíduo i, é a soma de uma constante, que é uma expectância dos escores observados, e de uma variável aleatória, que representa o erro de medida. Diferem na definição dessa constante: na teoria clássica, o escore verdadeiro é a expectância dos escores observados por todo o universo infinito de medidas repetidas do indivíduo i, enquanto na teoria da generalização o escore universal é a expectância dos escores observados do indivíduo i por todo o universo de generalização.

Em ambas, o erro de medida é concebido como a diferença entre o escore observado e uma constante, para cada indivíduo i – seja o escore verdadeiro, seja o escore universal. Apesar das duas compreenderem que são as mais diversas as fontes de erros de medida, divergem na definição do erro. Na teoria clássica, a definição de erro de medida está associada à concepção de medidas repetidas: o erro é

a diferença entre o escore observado em cada repetição da medida do indivíduo i e o escore verdadeiro. Na teoria da generalização, com a concepção do universo de observações admissíveis, o erro de medida é definido como a diferença entre o escore universal do indivíduo i e o escore observado em cada medida, em cada uma das infinitas condições possíveis que formam cada faceta incluída em cada estudo específico. Diferenças entre condições, ou facetas, passam a ser avaliadas explicitamente como efeitos de fontes de erros de medida.

Os processos de análise da variância têm sido empregados no estudo da confiabilidade dos escores observados no contexto das duas teorias. No caso da teoria clássica, os erros de medida são indivisíveis, embora se reconheça que diversas fontes são por eles responsáveis. Já na teoria da generalização os efeitos das várias fontes de variabilidade dos escores observados que são incluídas no estudo G ou D são explicitamente avaliados e sua inclusão no erro de medida é claramente determinada nos modelos e nos processos de análise da variância.

O conceito de confiabilidade é fundamentalmente o mesmo : tanto o coeficiente de fidedignidade como o de generalização representam a proporção da variância dos escores "constantes" (escores verdadeiros ou escores universais) na variância total dos escores observados na população de indivíduos submetidos a uma medida psicológica ou educacional. Obviamente, a divergência está na concepção de escores verdadeiros ou de escores universais, assim como na de erros de medida.

Os estudos G e D são uma extensão e uma especificação do processo de experimentações e de análises no desenvolvimento de medidas psicológicas e educacionais. As definições de tipos de erro Δ e δ, assim como a possibilidade de uso de diferentes modelos na análise dos componentes da variância abrem ao pesquisador alternativas para explorar as fontes de variabilidade no estudo de cada instrumento ou de cada processo de medida.

A teoria da generalização abre perspectivas para a investigação psicométrica, ao superar, por exemplo, o conceito de erro de medida indivisível, ou as dificuldades práticas do emprego de testes paralelos; entretanto, enfrenta em comum com a teoria clássica, críticas feitas tanto por psicometristas como por psicólogos cognitivistas. Entre outras, é o caso dos pressupostos que compõem modelos de escores observados e que são considerados pouco realistas. Apontam-se os

pressupostos de que erros de medida não são correlacionados, ou de que não variam conforme as medidas sejam de indivíduos que alcançam níveis de desempenho mais alto na escala ou daqueles de desempenho inferior – veja-se, a respeito, Snow e Lohman (1993) e Hambleton (1993).

Deve-se observar, entretanto, que psicometristas e psicólogos cognitivistas concordam em mostrar o caminho das pesquisas psicométricas, das pesquisas experimentais e do aprofundamento da fundamentação psicológica das medidas como a forma de procurar respostas para as questões que a teoria levanta.

De um lado, Lord (1980) explica que um modelo estatístico é expresso em termos matemáticos não definidos no "mundo real" – se o "mundo real" corresponde ao modelo é uma outra questão a ser respondida da melhor maneira possível pelos pesquisadores. De outro, Snow e Lohman (1993) acentuam que testes educacionais, fundamentados na teoria e na pesquisa empírica em psicologia, podem melhor contribuir para a compreensão dos fenômenos observados nas pesquisas de campo, ao mesmo tempo que o conhecimento das relações observadas no mundo empírico traz implicações que levam ao desenvolvimento teórico. Numa perspectiva mais prática, Linn (1993) chama a atenção para a urgência de combinarem esforços os psicólogos cognitivistas, os especialistas em medidas e os educadores para que as medidas psicológicas e educacionais possam contribuir de modo mais efetivo para facilitar a aprendizagem.

Referências bibliográficas

BENNETT, Randy E., ROCK, Donald A. *Generalizability, Validity, and Examinee Perceptions of a Computer-Delivered Formulating-Hypotheses Test. Educational Testing Service Research Report 93-4.* Princeton, N.J.: Educational Testing Service, 1993.

BENNETT, Randy E., MORLEY, Mary, QUARDT, Dennis, et al. *Psychometric and Cognitive Functioning of an Undertermined Computer-Based Response Type for Quantitative Reasoning. Educational Testing Service Research Report 98-29.* Princeton, N.J.: Educational Testing Service,1998.

BESSA, Nicia M. Fidedignidade de Notas Atribuídas a Redações: Enfoque Teórico e Empírico. *Educação e Seleção*. 1986, (14), p. 25-46.

BRENNAN, Robert L., GAO, Xiaohong, COLTON, Dean A. Generalizability Analyses of Work Keys Listening and Writing Tests. *Educational and Psychological Measurement*, 1995, 55(2), 157-176.

CRONBACH, Lee, GLESER, Goldine C., NANDA, Harinder, RAJARATNAM, Nageswari. *The Dependability of Behavioral Measurements : Theory of Generalizability for Scores and Profiles*. N.Y.: John Wiley,1972.

EBEL, Robert L. Estimation of the Reliability of Ratings. In: MEHRENS, W. A. e EBEL, R. L. (Eds.). *Principles of Educational and Psychological Measurement*. Chicago,Ill.: Rand McNally, 1967,116-131.

FELDT, Leonard S., BRENNAN, Robert L. Reliability. In: LINN, R.L. (Ed.), *Educational Measurement*. Phoenix, AZ : Oryx, 1993, 105-146.

GULLIKSEN, Harold, *Theory of mental tests*. N.Y.: John Wiley,1950.

HAMBLETON, Ronald K. *Principles and Selected Applications of Item Response Theory*. In: LINN, R.L. (Ed.), *Educational Measurement*. Phoenix, AZ : Oryx, 1993, 147-200.

LINN, Robert L. *Current Perspectives and Future Directions*. In: LINN, R.L. (Ed.), *Educational Measurement*. Phoenix, AZ : Oryx, 1993, 1-10.

LORD, Frederic M., NOVICK, Melvin R. *Statistical Theories of Mental Tests Scores*. Reading,Mass.: Addison-Wesley, 1968.

LORD, Frederic M. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, N. J.: Lawrence Erlbaum,1980.

MESSICK, Samuel. *Validity*. In: LINN, R.L. (Ed.), *Educational Measurement*. Phoenix, AZ : Oryx, 1993, 13-103.

SHAVELSON, Richard J., BAXTER, Gail P., GAO, Xiaohong. Sampling Variability of Performance Assessments. *Journal of Educational Measurement*, 1993, 30(3), 215-232.

SIRECI, Stephen G., WAINER, Howard, BRAUN, Henry, *Psychometrics*. Princeton, N. J.: Educational Testing Service, 1998.

SNOW,Richard E., Lohman,David F., *Implications of Cognitive Psychology for Educational Measurement*. In : Linn,R.L.(Ed.), *Educational Measurement*.Phoenix,AZ : Oryx,1993, 263-331.

STANLEY, Julian C. Reliability. In: THORNDIKE, R. L. (Ed.), *Educational Measuremen t(2nd.Ed.)*. Washington D. C.: American Council on Education, 1971, 356-442.

VIANNA, Heraldo M. Flutuações de Julgamentos em Provas de Redação. *Cadernos de Pesquisa*,1976, 19, 5-9.

WINER, B. J. *Statistical Principles in Experimental Design (2nd.Ed.)*. N. Y.: MmcGraw-Hill, 1971.