ÁREA TEMÁTICA
Ensino e Pesquisa em Administração e Contabilidade

Sobre Confiabilidade e Validade

Gilberto de Andrade Martins
Professor Titular do Departamento de Contabilidade da FEA/USP.
Coordenador do Curso de Bacharelado em Ciências Contábeis da FEA/USP. [martins@usp.br]

Recebido em 10/Agosto/2005
Aprovado em 20/Fevereiro/2006

Resumo

Para medir, avaliar ou quantificar informações financeiras, patrimoniais, de auditorias, arbitragens e controladoria, peculiares ao setor privado ou público, o profissional ou pesquisador precisará atentar para os critérios de significância e precisão dos instrumentos de medidas que irá utilizar: validade, ou validez e confiabilidade ou fidedignidade. O critério da validade diz respeito à capacidade do instrumento em medir de fato o que se propõe medir, enquanto a confiabilidade está relacionada com a constância dos resultados obtidos quando o mesmo indivíduo, ou objeto é avaliado, medido ou quantificado mais do que uma vez. Sem a devida atenção a essas características, as medidas coletadas, ou as aferições patrimoniais não serão merecedoras de crédito e de significância. Este artigo tem o objetivo de apresentar, explicar, exemplificar e discutir critérios para indicação do grau de confiabilidade: técnica do teste-reteste; técnica de formas equivalentes; metades partidas (*split-half*); confiabilidade a partir de avaliadores; coeficiente alfa de Cronbach, bem como técnicas para evidenciação da validade: validade aparente; de conteúdo; de critério; de constructo e validade total. São mostradas ilustrações dos critérios de avaliação e evidenciação da confiabilidade e validade nas Ciências Contábeis.

Palavras chave

Confiabilidade; Validade; Medidas; Ciências Contábeis; Avaliação.

Abstract

In order to assess, evaluate or quantify financial, equity, auditing and controllership oriented data related both to private and public sectors, the practioner or the researcher has to pay close attention to the significance and accurateness criteria of the research tools he is about to employ: validity and reliability. The validity criterion refers to the instrument capacity of assessing what it intends to assess; reliability deals with the constancy of results when the same individual or object is assessed, evaluated or quantified more than once. This article is aimed at explaining and discussing examples of proper criteria to indicate the reliability level: test-retest, equivalent form techniques, split-half, reliability based on evaluators, Cronbach´s alpha coefficient, as well as techniques towards validity evaluation: apparent validity, content validity, criterion validity, construct and total validity. Different illustrations of the criteria for assessing validity and reliability in the Accounting field are shown.

Key words

Reliability; Validity; Measures; Accounting; Assessment.

1. INTRODUÇÃO

O primeiro passo para elaboração de um instrumento de medidas é definir o que deve ser medido e como deve ser medido. Respostas a tais perguntas podem ser obtidas pela realização de pesquisa exploratória com objetivo de verificar os tipos de dados que realmente se referem à questão, ou constituem indicadores adequados da medida, bem como a melhor forma de obtê-los. A construção de qualquer instrumento de medidas – seja um questionário, um teste, ou outra técnica de aferição exige a observância de cuidados sem os quais não se poderá ter segurança quanto aos seus resultados. O sucesso de um instrumento de medidas é obtido quando se conseguem resultados merecedores de créditos para a solução de um problema de pesquisa ou relatório de trabalho profissional.

Neste artigo pretende-se apresentar, explicar, exemplificar e discutir critérios de exigências de

medidas provenientes de instrumentos de coleta de dados, e técnicas de aferição, para que se possa aceitá-los como geradores de boas medidas. Ainda que divergindo em alguns pontos, os autores são unânimes, em apontar dois critérios fundamentais de um bom instrumento de medidas: confiabilidade ou fidedignidade, e validade, ou validez. Registre-se também a pluralidade de nomes dados aos critérios de significação de medidas, daí um alerta ao leitor quando da análise e entendimento dessa matéria. Por exemplo, há autores que substituem a palavra confiabilidade por precisão, outros denominam consistente ao instrumento de medidas que neste texto denominamos confiável. (GRESSLER, 1989)

Toda medida deve reunir dois requisitos essenciais: confiabilidade e validade. Medidas confiáveis são replicáveis e consistentes, isto é, geram os mesmos resultados. Medidas válidas são representações precisas da característica que se pretende medir. Confiabilidade e validade são requisitos que se aplicam tanto às medidas derivadas de um teste, instrumento de coleta de dados, técnicas de aferição, quanto ao delineamento da investigação – a pesquisa propriamente dita. Neste artigo discute-se a validade e confiabilidade de medidas.

É comum apresentar-se a validade de um instrumento como o seu primeiro requisito, mas, considerando-se que para ser válida uma medida deve também ser confiável, não sendo verdadeira a recíproca, parece argumento razoável analisar-se a confiabilidade antes da validade. Em outras palavras, nem todo instrumento de medidas que apresenta confiabilidade tem validade, mas todo aquele que tem validade também apresenta confiabilidade. (SAMPIERI, 1996). Para ilustrar tal entendimento podemos analisar, por exemplo, o depoimento de uma testemunha: ela pode manter constante o seu depoimento, sem apresentar desvio do relato sobre o que ocorreu, isto é, ser confiável, mas isso não garante que o depoimento tenha validade, isto é, expresse o que de fato ocorreu. Por outro lado, se durante os depoimentos a testemunha não mantém constância na sua história, ou seja, não consegue apresentar confiabilidade nas suas explicações, poderemos concluir que o depoimento não é confiável, nem tão pouco apresenta validade.

2. CONFIABILIDADE

A confiabilidade de um instrumento para coleta de dados, teste, técnica de aferição é sua coerência, determinada através da constância dos resultados. Em outras palavras, a confiabilidade de uma medida é a confiança que a mesma inspira. Os instrumentos para medir fenômenos do mundo físico, em geral, oferecem um grau de confiança bastante elevado, devido à relativa estabilidade dos fenômenos observados. A comparação dos resultados de uma série de medidas de um elemento físico, em idênticas condições, fornece um elevado coeficiente de segurança, ou baixa margem de erro do aparelho de medição. Nem sempre o mesmo acontece em relação às medidas de variáveis do universo social onde a instabilidade dos fenômenos e fatos observados dificultam a própria construção de instrumentos de aferição, pois as contínuas modificações do ambiente tornam bem mais difíceis a determinação da constância das medidas, isto é, geralmente dificultam a obtenção de um elevado grau de confiabilidade. Ainda assim, a confiabilidade de um instrumento de medição de fenômenos sociais é obtida do mesmo modo: comparação dos resultados em situações semelhantes e sucessivas. Conforme explica Cozby (2003), confiabilidade de um instrumento de medição se refere ao grau em que sua repetida aplicação, ao mesmo sujeito ou objeto, produz resultados iguais. Por exemplo, ao se medir de forma constante a temperatura de uma sala climatizada, o termômetro que apresentar resultados diferentes em cada medição deve ser considerado não confiável, pois, sabemos que nessas condições, não há motivo para mudanças de temperatura. Se ocorrerem resultados alterados o instrumento de medidas não terá a característica de fidedignidade e seus resultados não serão confiáveis.

De maneira ampla, uma medida fidedigna é consistente e precisa porque fornece uma medida estável da variável. Em outras palavras, confiabilidade refere-se à consistência ou estabilidade de uma medida. Para facilitar a compreensão do conceito de confiabilidade de uma medida pode-se fazer analogia com o que se entende por um indivíduo confiável. Se você diz que alguém é confiável, provavelmente você quer dizer que a pessoa é fidedigna, consistente – se ela diz uma coisa hoje, dirá a mesma coisa amanhã. Se narrar a ocorrência de um acontecimento, manterá um relato consistente, não expressará versões do ocorrido. Um instrumento confiável também manterá 'a mesma história' em momentos distintos. Um exemplo corriqueiro pode nos ajudar a compreender ainda mais este conceito – diz-se que se tem um relógio confiável quando o instrumento nos fornece o tempo preciso, raramente adiantado ou atrasado. Segundo Selltiz (1987) uma medida confiável produzirá os mesmos resultados em sucessivas

aplicações sobre um mesmo sujeito ou objeto. Uma medida confiável não flutua entre uma leitura e outra do mesmo objeto ou sujeito. Se uma medida flutua entre uma e outra medição do mesmo objeto ou sujeito é porque há erro na mensuração. Entretanto, parte da flutuação deve ser entendida como resultante de diferenças reais entre medidas e parte representa erros de mensuração. O problema básico na avaliação dos resultados de qualquer mensuração é o de definir o que deve ser considerado como diferenças reais na característica medida, e o que deve ser considerado como variações devidas a erros de mensuração.

O desvio padrão (medida de dispersão em torno da média) pode ser um indicador do grau de confiabilidade de um instrumento de medidas. Assim é que: quanto menor o valor do desvio padrão maior será o grau de confiabilidade do instrumento de medidas. Além dessa maneira a confiabilidade de um instrumento de medidas pode ser determinada mediante diversas técnicas e procedimentos, sendo os mais conhecidos os seguintes:

2.1. Técnica do teste reteste

O instrumento de medidas é aplicado duas vezes a um mesmo grupo de pessoas, depois de um período de tempo entre as aplicações. Se a correlação entre os resultados das duas aplicações é fortemente positiva o instrumento pode ser considerado confiável. Quando a variável sob análise apresentar nível intervalar de mensuração, pode-se calcular o coeficiente de correlação linear de Pearson. (SAMPIERI, 1996).

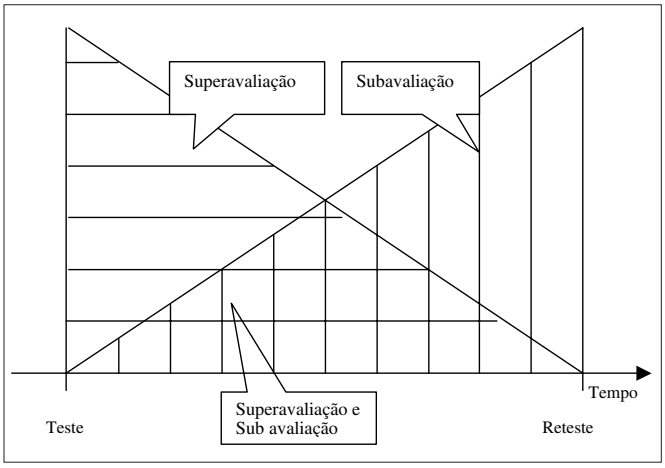
O período de tempo entre as medições é um fator a considerar quando da aplicação desta técnica. Períodos longos são suscetíveis às mudanças que podem comprometer a interpretação do coeficiente de confiabilidade obtido. Um tempo longo demais favorece a aquisição de novas aprendizagens. Se o período é curto, os resultados podem ser contaminados pelo efeito memória. No caso desta técnica o coeficiente de confiabilidade é também denominado coeficiente de estabilidade.

O intervalo longo entre o teste e reteste pode provocar uma sub avaliação da estabilidade. Tal conceito poderá ser melhor explicado através de um exemplo: vamos supor que foi aplicado um questionário com a seguinte pergunta: o que você prefere como sobremesa? Com as seguintes alternativas: (1) Sorvete, (2) Torta de morango e (3) Não sei.

Na primeira aplicação o respondente marcou a alternativa (3). Porém, as alternativas despertaram o respondente quanto às possibilidades de sobremesa. Depois de um longo tempo foi aplicado o reteste e ao deparar com a mesma questão a pessoa escolheu a alternativa (1). O pesquisador ao comparar as respostas pode ser induzido a afirmar que o instrumento de medidas não tem estabilidade, mas na verdade, trata-se de uma mudança real da pessoa. Esse efeito é chamado de sub avaliação da estabilidade.

O intervalo curto entre a aplicação de um teste e reteste também provoca um efeito conhecido como superavaliação da estabilidade. Esse efeito pode ser provocado pela lembrança das respostas que o indivíduo deu no primeiro teste e depois, simplesmente repete as respostas recordadas no reteste, ou seja, não são respostas espontâneas ou inteiramente pensadas. A **Ilustração 01** mostra os comportamentos desses efeitos.

Ilustração 1 – Comportamento dos efeitos de superavaliação e sub avaliação



Observa-se que, com o passar do tempo, o efeito da superavaliação diminui enquanto o efeito da subavaliação aumenta. O leitor poderá perguntar: qual é o pior dos dois efeitos? A superavaliação ou a sub avaliação da estabilidade? A resposta a essa questão deve ser analisada em termos da interpretação do coeficiente de estabilidade. O pesquisador estará mais seguro com uma interpretação dos efeitos da sub avaliação do que com a superavaliação, pois, no primeiro efeito o pesquisador concluirá que há necessidade de mais estudo sobre a medida em questão, enquanto o segundo efeito pode dar uma falsa segurança de estabilidade e se chegar a conclusões inválidas.

Conforme explicam Camines e Zeller (1979), para avaliar a confiabilidade pelo teste e reteste precisamos obter dois escores (medidas) de cada um de muitos indivíduos, ou objetos. Se a medida for confiável os dois escores, para cada indivíduo ou objeto, deverão ser muito semelhantes, e o coeficiente de correlação linear de Pearson positivamente elevado – acima de 85%. Este critério de avaliação da confiabilidade só poderá ser aplicado quando o nível de mensuração da variável é intervalar. Não é comum se ter duas medidas de uma variável para os mesmos indivíduos ou objetos, fato que limita a aplicação deste critério.

2.2. Técnicas de formas equivalentes

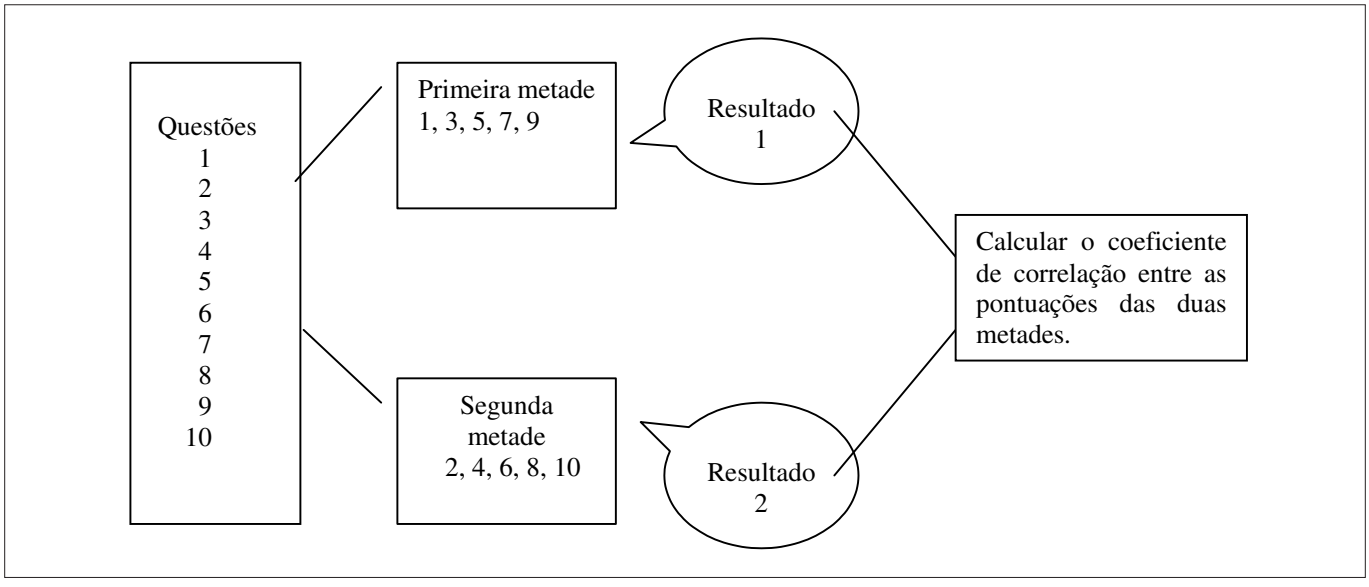
Neste procedimento não se aplica o mesmo instrumento de medidas às mesma pessoas ou objetos, mas duas ou mais versões equivalentes do instrumento de medidas. As versões são similares em conteúdo, instruções e demais características. As versões – geralmente duas – são administradas a um mesmo grupo de indivíduos dentro de um período relativamente curto. O instrumento é confiável se a correlação entre os resultados das duas aplicações é fortemente positiva, ou seja, os padrões de respostas devem variar pouco entre as aplicações. A maior limitação de aplicação desta técnica é que nem sempre se dispõe de duas formas distintas de um instrumento de medidas com iguais objetivos (ANASTASI,1965).

2.3. Técnicas das metades partidas (*split-half*)

Contrariamente às técnicas anteriores este procedimento requer apenas uma aplicação, ou seja, consiste em avaliar a confiabilidade usando respostas obtidas em uma única aplicação do instrumento de medidas.

Para um melhor entendimento sobre a técnica das Metades Partidas (*split-half*) vamos considerar a seguinte configuração: apresentamos aos respondentes um instrumento de medidas com 10 questões, tais que as questões 1 e 2 são equivalentes em conteúdo e dificuldade, o mesmo raciocínio serve para as questões 3 e 4 e assim por diante. O resultado dessa divisão é que temos um conjunto de questões (1,3,5,7,9) equivalente, em termos de conteúdos e dificuldades, ao conjunto de questões (2,4,6,8,10). Segundo Sampieri (1996), o conjunto de todas as questões do teste é dividido em duas metades e as pontuações, ou resultados, de ambas são comparados. A comparação é feita através do cálculo do coeficiente de correlação linear de Pearson entre o escore total de cada indivíduo na primeira metade do teste e o escore total na segunda metade do teste. Se o instrumento é confiável as pontuações das duas metades devem estar fortemente relacionadas. Em outras palavras, um indivíduo, com baixa pontuação em uma das metades, tenderá a ter também uma baixa pontuação na outra metade. Quanto mais semelhantes forem os escores das duas metades, maior será a correlação e mais confiável o instrumento. A confiabilidade calculada dessa maneira

Ilustração 2 – A Técnica das Metades Partidas (Split-half).



é interpretada, por alguns autores, como indicador de consistência interna.

Ainda, conforme Sampieri (1996) a confiabilidade varia de acordo com o número de ítens do instrumento de medição. Quanto mais itens maior a possibilidade de se avaliar a confiabilidade do instrumento. Alternativamente, pesquisadores contrapõem as questões impares com as pares. É preciso que os totais de escores sejam variáveis com níveis de mensuração intervalar. A **Ilustração 02** mostra a prática desta técnica.

2.4. Confiabilidade a partir de avaliadores

Há situações de pesquisa em que diferentes avaliadores observam comportamentos e fazem medições ou julgamentos. Se dois avaliadores (juízes) observarem o mesmo comportamento, a partir das mesmas instruções e igual treinamento, a confiabilidade das medidas será dada pelo cálculo do coeficiente de correlação linear de Pearson entre os escores dos dois juizes. Para tratamento de variáveis com níveis de mensuração ordinais (quando os juizes classificam ou colocam em ordem) podem ser calculados os coeficientes de Spearman ou de Kendall.

2.5. Coeficiente alfa de Cronbach

Conforme explicam Carmines e Zeller (1979), este Coeficiente foi desenvolvido por J.L. Cronbach, e o seu cálculo (), alfa, carece de uma única aplicação do instrumento de medição, produzindo valores entre 0 e 1, ou entre 0 e 100%. Quando > 70% diz-se que há confiabilidade das medidas. A expressão do coeficiente é dada por:

$$\alpha = \frac{N\overline{\rho}}{[1+\overline{\rho}(N-1)]}$$

Onde:

N = número de itens;

$\overline{\rho}$ = média dos coeficientes de correlação linear (Pearson) entre os itens.

$0 <= \alpha <= 1$ ou $0 <= \alpha <= 100\%$.

São calculadas todas as correlações (ρ) entre o escore de cada item e o escore total dos demais itens. O valor de alfa é a média de todos os coeficientes de correlação. As correlações item-total e o valor do alfa de Cronbach são reveladoras porque fornecem informações sobre cada item individual. Itens que não estão correlacionados com os demais podem ser eliminados da medida para aumentar a confiabilidade.

2.6. Coeficiente KR-20

Com finalidade semelhante ao coeficiente de Cronbach este indicador do grau de confiabilidade de um instrumento de medição foi desenvolvido por Kuder e Richardson (1937). É utilizado quando os testes têm respostas dicotômicas: sim/não; 0/1 etc.(CARMINES e ZELLER, 1979).

3. VALIDADE

Em termos gerais a validade se refere ao grau em que um instrumento realmente mede a variável que pretende medir. Em outras palavras, um instrumento é válido na extensão em que mede aquilo que se propõe medir. Por exemplo, um instrumento válido para medir a capacidade de leitura deve medir realmente essa característica e não outras características, como por exemplo, conhecimento prévio. Para facilitar a compreensão desse conceito vamos supor que estamos interessados em medir a capacidade de leitura de uma pessoa e para isso aplicamos um teste simples que se resume em ler a história dos três porquinhos e depois a pessoa nos conta o que leu. Será que esse teste mede o que realmente se propõe medir? Não, necessariamente, pois podemos ter ocorrência em que uma pessoa, que não sabe ler, sair-se bem no teste porque já ouviu essa história antes, ou seja, para essa pessoa, o teste não mediu a capacidade de leitura e sim o conhecimento prévio. Outro exemplo: quando estudantes brasileiros fazem um teste de QI (Quociente de Inteligência) em inglês, tal teste é muito mais uma medida da proficiência na língua inglesa do que uma medida (válida) de inteligência, pois podemos ter estudantes brasileiros inteligentes, mas que foram mal no teste por não compreenderem a língua inglesa.

Conforme lembra Gressler (1989) a questão fundamental para se admitir a validade de um instrumento de medidas é dada pela resposta à seguinte pergunta: Será que se está medindo o que se crê que deve ser medido? Se a resposta é sim, sua medida é válida, se não, não é.

A validade é um critério de significância de um instrumento de medidas com diferentes tipos de evidências: validade aparente, validade de conteúdo, vali-

dade de critério e validade de construto (MORON, 1998). A validade da medida depende da adequação do instrumento em relação aquilo que se quer medir. Ou seja, a adequação do instrumento dependerá do uso que dele se fizer. Por exemplo, existem vários instrumentos para medir o tempo: desde a posição do sol, relógio de areia, relógios que marcam horas, minutos e segundos, até aqueles mais precisos que determinam frações de segundos. Pois bem, a utilização de um ou outro desses instrumentos dependerá do que vai se medir. Um jogo de futebol requererá um relógio que assinale até segundos, não sendo suficiente a posição do sol para determinar o término da partida. Por outro lado, o controle de uma corrida de cavalos exigirá um instrumento mais preciso, como o cronômetro. Porém o lavrador do campo saberá quando é hora de almoço, ou quando seu dia de trabalho termina, pela simples posição do sol. Ou seja, a validade de uma medida nunca é absoluta, mas sempre relativa – um instrumento de medidas não é simplesmente válido, porém, será válido para este ou aquele objetivo. Não há validade em termos gerais.

3.1. Validade Aparente

A técnica mais simples, porém menos satisfatória, para avaliar a validade é denominada validade aparente, que nos indica se a medida aparentemente mede aquilo que pretende, como explica (GIL, 1999). A validade aparente não é sofisticada, avalia apenas, considerando a definição teórica de uma variável, se a medida parece, de fato, medir a variável sob estudo. Isto é, o procedimento usado para medir a variável parece ser uma definição operacional correta da variável teórica? Validade aparente é avaliada por um juiz, ou grupo de juízes, que examinam uma técnica de mensuração e decidem se ela mede o que seu nome sugere. A avaliação da validade aparente é um processo subjetivo. Todo instrumento deve passar pela avaliação da validade aparente. Todo pesquisador que escolhe, ou constrói, um instrumento de medidas é um juiz que decide se o instrumento de fato mede a variável que ele deseja estudar. A validade aparente não basta para se concluir se uma medida é de fato válida, todavia sem algum indicador positivo de validade aparente não terá sentido avaliações dos outros critérios de validade.

A validade aparente refere-se ao fato do instrumento de medidas parecer válido, ou não, aos sujeitos, ao pessoal administrativo que decide quanto ao seu emprego, e a outros observadores não treinados tecnicamente. À primeira vista o leitor poderá concluir que a validade aparente não tem muita importância e utilidade pois lhe falta uma construção mais técnica. No entanto, a validade aparente é uma característica necessária porque se o instrumento de medidas parece, aos olhos dos respondentes, irrelevante, inadequado, tolo ou infantil, a falta de validade aparente poderá comprometer todo o estudo. Tal situação pode ser observada, por exemplo, em testes que inicialmente foram desenvolvidos para crianças e que depois foram também aplicados para adultos. Esses testes enfrentaram sérias resistências e críticas dos adultos por falta de validade aparente, pois, para adultos pareciam irrelevantes, inadequados e infantis.

3.2. Validade de conteúdo – Evidências relacionadas ao conteúdo

Segundo Sampieri (1996), a validade de conteúdo se refere ao grau em que um instrumento evidencie um domínio específico de conteúdo do que pretende medir. É o grau em que a medição representa o conceito que se pretende medir. Por exemplo, uma prova de operações aritméticas não terá validade de conteúdo se incluir somente problemas de adição e excluir problemas de subtração, multiplicação e divisão. Um instrumento de medição deve conter todos os itens do domínio do conteúdo das variáveis que pretende medir. Assim, pode parecer que uma simples verificação do conteúdo do teste é suficiente para estabelecer a validade com relação a esse objetivo, no entanto, a solução não é tão simples. Uma dificuldade é apresentada pelo problema da amostragem do conteúdo. A área de conteúdo a ser testada precisa ser sistematicamente analisada a fim de se assegurar que todos os aspectos fundamentais sejam, adequadamente, e em proporções corretas, abrangidos pelos itens do teste. Para se ter maior garantia da validade de conteúdo de um instrumento de medidas, a área de abrangência do conteúdo deve ser inteiramente descrita antes, e não depois da construção de um do teste, ou qualquer outro instrumento de coleta de dados.

3.3. Validade de Critério – Evidências relacionadas a um critério

Conforme Kaplan (1975), a validade de critério estabelece a validade de um instrumento de medição

comparando-o com algum critério externo. Este critério é um padrão com o qual se julga a validade do instrumento. Quanto mais os resultados do instrumento de medidas se relacionam com o padrão (critério) maior a validade de critério. Se o critério se fixa no presente, temos a validade convergente – os resultados do instrumento se correlacionam com o critério no mesmo momento ou ponto no tempo. Por exemplo, um roteiro de entrevista para levantar as preferências eleitorais pode ser validado comparando-se os resultados da pesquisa com os resultados da eleição. Assim, quanto mais próximos os resultados da pesquisa dos resultados das eleições, maior o grau de validade convergente do instrumento de coleta de dados. Se o critério se fixa no futuro temos a validade preditiva.Segundo Sampieri (1996), validade para predizer refere-se à extensão a qual o instrumento (geralmente teste) prediz futuros desempenhos de indivíduos. Um teste tem validade para predizer quando efetivamente indica como o objeto em estudo desenvolverá no futuro uma outra tarefa ou incumbência. A validade peditiva é muito importante para testes que são usados com propósitos de selecionar e classificar candidatos a concursos para admissão, exames vestibulares etc. Conforme já explicado a validade de predizer é estabelecida através de correlações dos resultados do teste com subseqüente medida de um critério. A identificação de uma medida critério que se adeqüe ao instrumento que está sob avaliação, geralmente, constitui desafio ao investigador. Por exemplo, um teste para determinar a capacidade administrativa de altos executivos pode ter validade preditiva comparando-se os resultados do teste com o futuro desempenho dos executivos avaliados pelo referido instrumento. Além disso, o instrumento de medidas não deve estar relacionado a variáveis que não lhe dizem respeito, ou seja, com um falso critério. Essa característica é formalmente conhecida como validade discriminante.

A comparação entre os resultados (medições) de um instrumento com outro critério exterior é também chamada de validade empírica. Conforme Cozby (2003), quando um teste, ou instrumento, consegue distinguir indivíduos sabidamente diferentes, diz-se que o teste, ou instrumento de medidas apresenta validade simultânea. Por exemplo, se você estivesse desenvolvendo um teste para medir o nível de consciência política dos indivíduos e conseguisse distinguir, pelo teste, os 'sabidamente de esquerda' dos 'sabidamente de direita', seu teste teria validade simultânea, pois além de medir o grau de consciência política também conseguiria distinguir os indivíduos de esquerda e de direita.

A distinção lógica entre validade de predição e validade simultânea baseia-se não no tempo, mas nos objetivos da aplicação. A validade simultânea é significativa para testes empregados para o diagnóstico de situação existente, e não para a predição de resultados futuros.

Como o critério para a validade simultânea sempre existe no momento da aplicação, poder-se-ia perguntar: qual a função da aplicação em tais situações? Basicamente, esses testes apresentam um substituto mais simples, mais rápido ou menos dispendioso do que os dados do critério. Por exemplo, se o critério para se concluir se um indivíduo é neurótico consiste na observação contínua de um paciente, durante um período de duas semanas de hospitalização, um teste capaz de selecionar os neuróticos, dentre os casos duvidosos, reduziria consideravelmente o número de pessoas que exigiriam essa observação extensiva.

3.4. Validade de Constructo – Evidências Relacionadas ao Constructo

Um constructo, ou uma construção, é uma variável, ou conjunto de variáveis, isto é, uma definição operacional robusta que busca representar o verdadeiro significado teórico de um conceito. Conforme explica Gressler (1989), a validade de constructo será dada pela resposta à questão: em que medida a definição operacional (constructo) de um conceito de fato reflete seu verdadeiro significado teórico?

A validade de constructo se refere ao grau em que um instrumento de medidas se relacione consistentemente com outras medições assemelhadas derivadas da mesma teoria e conceitos que estão sendo medidos.

Segundo Sampieri (1996), dificilmente a validade de constructo será estabelecida em um único estudo. Ela é construída por vários estudos que investigam a teoria do constructo particular que está sendo medido. Medidas de variáveis do campo das ciências sociais aplicadas têm 'vida limitada'. Com o acúmulo de resultados de pesquisas, os investigadores descobrem limitações e criam novas medidas para corrigir possíveis problemas. Esse processo leva ao aprimoramento das medidas e a uma compreensão mais completa das variáveis subjacentes que estão sendo estudadas.

Ainda, conforme Gressler (1989), no caso de testes da área educacional, a validade curricular refere-se à extensão em que a amostra representada nas

questões do teste – constructo – abrange a matéria lecionada, ou todos os conteúdos curriculares. O processo de validação de um constructo deve, necessariamente, estar vinculado a uma teoria. Não é possível levar a cabo uma validação de constructo, a menos que exista um marco teórico que suporte o constructo em relação a outras definições.

3.5. Validade Total

A validade total, segundo Sampieri (1996) é obtida pela soma das validade de conteúdo, de critério e de constructo. Assim, a validade de um instrumento de medição se verifica com base nessas três evidências. Quanto mais evidências de validade de conteúdo, validade de critério e validade de constructo de um instrumento de medidas, maiores são as evidências que, de fato, está se medindo o que se pretende medir.

Como já foi explicado, um instrumento de medição pode ser confiável (apresenta confiabilidade) e não, necessariamente, ser válido. Um instrumento pode ser consistente nos resultados que produz, porém não medir aquilo que pretende. Ou seja, um instrumento de medição para, de fato, representar a realidade deve ser confiável e válido.

3.5.1. *Configuração – Validade Total – Avaliação de Conhecimentos sobre Contabilidade*

Um instrumento de medidas tem validade quando mede o que realmente se propõe medir e, conforme exposto neste texto, há várias formas de evidenciar a validade que são: aparente, de conteúdo, de critério e de constructo. Para exemplificar os critérios de evidenciação da validade será usado um instrumento de medidas bastante conhecido por todos: uma prova para avaliação do aprendizado sobre Contabilidade. A prova contém as seguintes questões:

(1) Cite as diferenças entre o custeio direto e indireto.
(2) O que é ponto de equilíbrio?
(3) O que é margem de contribuição?

Com esse instrumento nos propomos medir se a pessoa conhece, ou não, Contabilidade. Ao analisarmos as perguntas percebemos que as questões abordadas referem-se à Contabilidade, portanto, essa prova tem validade aparente porque, aparentemente, mede características que podem indicar se uma pessoa conhece, ou não, Contabilidade. O fato da prova apresentar validade aparente não significa que é valida pois essa evidência de validade é muito frágil.

Ao analisar o conteúdo da prova, nota-se que as questões tratam apenas de uma parte da Contabilidade, ou seja, o conteúdo é insuficiente para medir se uma pessoa conhece, ou não, essa disciplina. Como o conteúdo da prova não é suficientemente abrangente para qualificar a característica pesquisada, essa prova não tem validade de conteúdo para o objetivo proposto. Ressaltamos, mais uma vez, que a validade de um instrumento não é absoluta e sim relativa, ou seja, essa prova não tem validade de conteúdo para o propósito a que se refere: avaliar se o respondente conhece, ou não, Contabilidade. Porém, pode vir a apresentar essa modalidade de validade se estivéssemos interessados em qualificar se uma pessoa tem, ou não, conhecimento básico sobre Contabilidade de Custo. Para continuar com nossa analogia vamos fazer o seguinte raciocínio: se esse teste tem capacidade de distinguir entre indivíduos sabidamente distintos: as pessoas que dominam e as pessoas que não dominam Contabilidade, logo esse teste tem validade simultânea e validade discriminante.

Admitindo-se que qualquer pessoa que acertar mais de 90% deste teste será aprovada no Exame de Suficiência do Conselho Federal de Contabilidade, poderemos afirmar que o teste apresenta validade preditiva, pois tem capacidade de identificar diferenças futuras: passar, ou não passar, no Exame de Suficiência.

Por outro lado, se considerarmos que qualquer pessoa que acertar mais de 95% do teste será qualificada como alguém com QI elevado, o teste também terá validade de critério, pois apresenta uma forte relação com um indicador de inteligência.

4. UMA APLICAÇÃO DOS CONCEITOS DE CONFIABILIDADE E VALIDADE NAS CIÊNCIAS CONTÁBEIS

Antes das considerações finais vamos propor algumas reflexões sobre os conceitos de validade e de confiabilidade nas Ciências Contábeis. Mais especificamente gostaríamos de buscar uma resposta à questão: de que forma esses conceitos, difundidos na metodologia de pesquisa, podem auxiliar o contador?

Para responder a essa questão vamos fazer uma analogia dos conceitos discutidos com o mundo contábil, ou seja, deixar de aplicar o conceito de validade e de confiabilidade somente nos instrumentos de coleta de dados, e responder a pergunta: Quais são os principais instrumentos de coleta de dados de um contador?

De acordo com Moron (1998): "Os instrumentos de coleta de dados têm a função de ligar o que o pesquisador quer saber com a realidade, ou seja, os instrumentos de pesquisa são utilizados para ler a realidade". Levando-se em consideração esse raciocínio podemos afirmar que os instrumentos de coleta de dados utilizados pelos contadores são as demonstrações financeiras como balanços, demonstrativos de resultados etc., porque através delas os contadores transmitem as realidades das empresas para o mercado.

De acordo com Hendrisken (1999), com base nas hipóteses de mercado eficiente, pesquisas empíricas confirmam a visão de que o lucro contábil possui conteúdo informacional tanto que o mercado continua a exigir a sua mensuração e publicação. Essa opinião é reforçada pelo SFAC 1 que diz "a principal preocupação da divulgação financeira é o fornecimento de informações sobre o desempenho de uma empresa, com base em medidas de lucro e seus componentes".

Sabemos que o lucro contábil é apurado de acordo com os princípios contábeis, oferecendo, dessa maneira, condições para se afirmar um elevado grau de confiabilidade. Se vários contadores trabalharem, independentemente, com os mesmos números, devem chegar a resultados semelhantes, ou seja, o lucro contábil, determinado dessa maneira, é confiável. Mas, será que o lucro contábil é valido?

Vamos imaginar uma situação onde temos duas empresas A e B. Ambas investiram em ações, sendo que a empresa A comprou ações X, enquanto a empresa B comprou ações Y. Além disso, as duas empresas compraram a mesma quantidade e a cotação das duas ações eram iguais na data de compra. Suponhamos que a quantidade comprada foi um lote de 1.000 ações cotado à $ 500, então, temos as seguintes situações patrimoniais:

Empresa B

Ativo	Passivo
Ações X 1.000	Lucro 500 Capital 500

Empresa A

Ativo	Passivo
Ações X 500 Capital 500	

No exercício seguinte, tanto as ações X como Y valorizaram 100%, portanto, a cotação atingiu $ 1.000 por lote de mil. Entretanto, a empresa A não realizou nenhum tipo de operação, enquanto a empresa B vendeu suas ações, e em seguida, comprou ações da empresa X, mesmas ações mantidas pela empresa A.

Nesta situação, o lucro orientado pelos princípios contábeis, mostra a seguinte situação financeira das duas empresas:

Empresa B

Ativo	Passivo
Caixa 1.000	Lucro 500 Capital 500

Empresa B

D.E.R	
Venda	1.000
Custo	(500)
Lucro	500

Empresa B

Ativo	Passivo
Ações X 1.000	Lucro 500 Capital 500

Empresa A

Ativo	Passivo
Ações X 500	Capital 500

Na essência econômica as duas empresas podem ser consideradas exatamente iguais, pois possuem o mesmo ativo. No entanto, a contabilidade oferece uma visão que induz ao usuário da informação uma conclusão errada de que a empresa B tem mais riqueza que a empresa A. É interessante notar que ao se elaborar tais demonstrações financeiras os princípios contábeis foram respeitados, tais como: da realização, do custo histórico como base de valor e da confrontação entre receitas e despesas.

Esses princípios são dotados de objetividade e conservadorismo, pois o custo histórico foi comprovado a partir da nota fiscal e o reconhecimento da receita na transferência de propriedade. Não reconhecer o aumento da riqueza pela simples variação do ativo no mercado é uma postura conservadora.

O lucro contábil apurado de acordo com esses princípios é extremamente objetivo e conservador e, muitas vezes, não reflete a realidade econômico-financeira da empresa. A partir deste simples exemplo pode-se notar que a medição da riqueza, através do lucro contábil, pode não ser válida, ou seja, dessa forma, dependendo do propósito que se deseja, não se está medindo aquilo que se pretende medir.

Existe um consenso quanto ao lançamento de ativos pelo seu valor de aquisição no momento da compra, no entanto, a discordância nasce em torno de qual valor deve ser usado até a sua baixa. Para exemplificar, poderemos raciocinar do seguinte modo: no momento da aquisição as mercadorias são contabilizadas pelo seu custo corrente que com o passar do tempo torna-se custo histórico e no momento da venda as mercadorias são reavaliadas ao preço de venda, porém, em outra conta que pode ser contas a receber ou caixa. Percebe-se que a discussão não é em torno de qual é a medida a ser usada (se de aquisição ou de venda), mas quando usá-la, ou em outras palavras, a questão se resume em quando deve ser feita essa reavaliação. Alguns defendem a não realização de qualquer reavaliação até o momento da venda, enquanto outros defendem a marcação do ativo ao mercado, trazendo volatilidade aos demonstrativos financeiros.

A Contabilidade baseada nos custos históricos tem vantagens como a ausência de viés no procedimento da reavaliação, certeza relativa sobre a conversão esperada em dinheiro e capacidade de medir as despesas associadas. No entanto, ao se adotar o custo histórico como métrica compromete-se a validade da medida, sobretudo, para os ativos negociados em um mercado firme e organizado.

5. UMA CONFIGURAÇÃO CONFIABILIDADE E VALIDADE DE UMA ESCALA DE ATITUDE

Recente estudo (Giraldi et all, 2005) desenvolveu pesquisa para levantar a atitude de um segmento de consumidores estrangeiros em relação aos calçados brasileiros. Lembram que atitude é uma predisposição aprendida para um comportamento consistentemente favorável ou desfavorável em relação a um determinado objeto. Para compreender a relação entre atitude e comportamento são elaborados modelos que capturam dimensões subjacentes de uma atitude a fim de melhor explicar ou prever comportamentos, no caso, de consumidores. Dentre os modelos escolheram o de atitude de três componentes. O componente cognitivo consiste nas cognições do indivíduo, ou seja, o conhecimento e as percepções que foram adquiridos pela combinação entre experiência direta com o objeto de atitude e as informações de várias fontes. O componente afetivo representa as emoções ou sentimentos dos consumidores em relação a um produto ou marca em particular. Enquanto o componente conativo relaciona-se com a probabilidade com que um indivíduo irá adotar um comportamento específico diante do objeto de atitude.

É tarefa deveras complexa e difícil medir construtos dessa natureza – comuns nos estudos comportamentais – pois uma atitude é um construto que existe na mente dos indivíduos, não podendo ser observada diretamente, como o peso ou a altura de uma pessoa. Para tanto são utilizadas escalas, geralmente do tipo Likert, onde o respondente escolhe o ponto que melhor expressa seu entendimento em relação à variável que está sendo medida. Na investigação sob análise foram utilizados escalas com cinco pontos, orientados por concordo totalmente até discordo totalmente, para as seguintes dimensões:

Componentes da atitude	Afirmações
Cognitivo	Os calçados brasileiros possuem boa reputação
	Os calçados brasileiros são caros
	Os calçados brasileiros têm prestígio
	Os calçados brasileiros são de alta qualidade
Afetivo	Eu gosto dos calçados brasileiros
	Eu acho os calçados brasileiros melhores do que os de outros países
	Eu admiro os calçados brasileiros
	Eu tenho simpatia pelos calçados brasileiros
Conativo	Eu compraria calçados brasileiros
	Eu recomendaria calçado brasileiro a um amigo
	Eu prefiro calçado brasileiro a calçados de outros países

Cada ponto da escala tem um valor – no caso de 1 a 5. Avaliações da confiabilidade e da validade dessa medida de atitudes poderiam ser realizadas da seguinte maneira.

Quanto a confiabilidade avaliada pela técnica do Teste-Reteste, teríamos que calcular o coeficiente de correlação entre as notas atribuídas pelos respondentes em duas épocas suficientemente distantes para se evitar efeitos memória. Se as associações entre as duas notas forem expressivas, poderemos afirmar que essa escala de medida da atitude em relação aos calçados brasileiros é confiável.

Se os autores do referido estudo pudessem aplicar, ao mesmo grupo de respondentes, uma outra versão do escalonamento utilizado na primeira aplicação, poderíamos dizer que o instrumento apresenta forte grau de confiabilidade se a correlação entre os resultados das duas aplicações for expressivamente positivo. Nesta situação teríamos a aplicação da técnica de formas equivalentes para se avaliar a confiabilidade.

Na configuração que estamos analisando a prática da técnica das metades partidas (*split-half*) poderia ser aplicada calculando-se a correlação entre os escores (soma dos valores) das duas metades de questões formuladas, por exemplo, pelas questões ímpares em um grupo e pares em outra metade. Se a associação entre os escores for expressiva, poderemos dizer que a escala de medidas tem confiabilidade.

Ainda em relação a avaliação da confiabilidade poderíamos calcular o coeficiente de Cronbach. Se o coeficiente for superior a 0,70, poderemos afirmar a confiabilidade da escala.

Para se ter indicações de que o escalonamento construído pelos autores mede atitude em relação aos calçados brasileiros precisamos avaliar a validez do instrumento. A validade aparente – a medida mede aquilo que pretende medir? – foi garantida, vez que os autores se apoiaram em estudos assemelhados para a construção da escala utilizada. Isto é: aparentemente o conjunto das afirmações avaliadas pelos respondentes mede a atitude desejada.

O aproveitamento de um construto utilizado por outros pesquisadores oferece garantias de validade de conteúdo. O construto formado pelos três componentes já havia sido utilizado por outros pesquisadores, condição necessária para se dizer que o instrumento apresenta validade de conteúdo. Para se avaliar a validade de critério precisaríamos comparar os resultados obtidos pela aplicação deste instrumento com resultados alcançados por outro instrumento já testado – confiável e válido – que medisse atitude em relação aos calçados brasileiros. Quanto mais próximos fossem os resultados mais elementos teríamos para avaliar a validade de critério. Por outro lado a validade de construto poderá ser aferida por evidências de que de fato o construto – atitude composta por três componentes – reflete o verdadeiro significado teórico de, no caso, medir atitude em relação a um produto: calçados brasileiros.

6. CONSIDERAÇÕES FINAIS

Já disse Hegel: medida é uma síntese da qualidade e da quantidade. Medir é determinar, tendo por base uma escala fixa, um padrão de unidade, e uma grandeza. Para medir, avaliar ou quantificar informações financeiras, patrimoniais, de auditorias, arbitragens e controladoria, peculiares ao setor privado ou público, o profissional ou pesquisador precisará atentar para os critérios de significância e precisão dos instrumentos de medidas que irá utilizar: validade, ou validez e confiabilidade ou fidedignidade.

Este texto apresentou, buscou explicações e mostrou exemplos sobre os critérios de exigências de medidas provenientes de testes, instrumentos de coleta de dados, e técnicas de aferição, para que se possa aceitá-los como geradores de boas avaliações. Medidas confiáveis são replicáveis e consistentes, isto é, geram os mesmos resultados em sucessivas medições, enquanto medidas válidas são representações precisas da característica que se pretende medir. A partir de um exemplo mostrou-se que nem todo instrumento de medidas que apresenta confiabilidade tem validade, mas todo aquele que tem validade também apresenta confiabilidade. Para avaliação da confiabilidade discutiu-se a técnica do Teste-Reteste: o instrumento de medidas é aplicado duas vezes a um mesmo grupo de pessoas ou objetos, depois de um período de tempo entre as aplicações. Se a correlação entre os resultados das duas aplicações é fortemente positiva o instrumento pode ser considerado confiável. Outro procedimento apresentado foi o uso de formas equivalentes, isto é, critérios assemelhados de aferição são aplicados aos mesmos elementos. A forte correlação entre os resultados do instrumento que se pretende utilizar e um outro assemelhado indicará elevado grau de confiabilidade. A técnica das Metades Partidas (*split-half*) determina que o conjunto de todas as questões do teste seja dividido em duas metades e as pontuações das metades sejam comparadas. A comparação é feita através do cálculo do coeficiente de correlação linear de Pearson entre o escore total de cada indivíduo na

primeira metade do teste e o escore total na segunda metade do teste. Se o instrumento é confiável as pontuações das duas metades devem estar fortemente relacionadas. A confiabilidade poderá também ser avaliada por dois juizes que observam o mesmo comportamento, a partir das mesmas instruções e igual treinamento. A confiabilidade é dada pelo grau de correlação entre os dois avaliadores. Além disso foram apresentados os coeficientes de Cronbach e o KR-20. Quanto aos critérios de aferição da validade vimos as seguintes: validade aparente – técnica simples, menos satisfatória, que nos indica se a medida, aparentemente, mede aquilo que pretende medir. A validade de conteúdo se refere ao grau em que um instrumento evidencie um domínio específico de conteúdo do que pretende medir. É o grau em que a medição representa o conceito que se deseja mensurar. A validade de critério estabelece a validade de um instrumento de medição comparando-o com algum critério externo. Este critério é um padrão com o qual se julga a validade do instrumento. Quanto mais os resultados do instrumento de medidas se relacionam com o padrão, maior a validade de critério. Por outro lado a validade de constructo se refere ao grau em que um instrumento de medidas se relacione consistentemente com outras medições assemelhadas derivadas da mesma teoria e conceitos que estão sendo medidos. Assim é que a validade total é obtida pela soma das validades de conteúdo, de critério e de constructo.

Por último foram apresentadas e discutidas configurações que ilustram aplicações dos critérios de confiabilidade e validade para uma situação do mundo contábil e aplicação em um escala de atitude.

7. REFERÊNCIAS

ANASTASI, Anne. **Teste Psicológicos: teoria e aplicação**. São Paulo: EDUSP, 1965.

CARMINES, Eduard. G. &ZELLER, Richard.A. **Reliability and Validity Assessment**. 3a ed., USA, Sage Publications, 1979.

COZBY, Paul C. **Métodos de pesquisa em ciências do comportamento**. São Paulo: Atlas, 2003.

GIL, Antonio Carlos. **Pesquisa Social.** 5a ed. São Paulo: Ed. Atlas, 1999.

GIRALDI, J.M.E. et all. Atitude de consumidores estrangeiros com relação a produtos brasileiros: Uma investigação do setor calçadista no Brasil. **Revista de Gestão USP**. São Paulo: v.12, n.3, p. 75-90, julho/setembro 2005.

GRESSLER, Lori Alice. **Pesquisa educacional.** São Paulo: Loyola, 1989.

HENDRIKSEN, Eldon S. BREDA, Michael F. Van. **Teoria da Contabilidade**. São Paulo: Atlas, 1999.

KAPLAN, Abraham. **A conduta na pesquisa**. São Paulo: EDUSP,1975.

MORON, Marie Anne Macadar. Dissertação: **Concepção, Desenvolvimento e Validação de Instrumentos de Coleta de Dados para Estudar a Percepção do Processo Decisório e as Diferenças Culturais.** Porto Alegre, 1998.

SAMPIERI, Roberto Hernández. COLLADO, Carlos Fernández. LUCIO, Pilar Baptista. **Metodología de la investigación.** México: McGRAW HILL, 1996.

SELLTIZ, Claire. WRIGHTSMAN, Lawrence Samuel. COOK, Stuart Wellford. KIDDER, Louise H. **Métodos de pesquisa nas relações sociais.** São Paulo: EPU, 1987.