Psicologia: Teoria e Pesquisa
Abr-Jun 2016, Vol. 32 n. 2, pp. 1-7

e322225
doi: http://dx.doi.org/10.1590/0102-3772e322225

Artigos Originais

Variância Média Extraída e Confiabilidade Composta: Indicadores de Precisão[1] [2]

Felipe Valentini[3]
Universidade Salgado de Oliveira
Bruno Figueiredo Damásio
Universidade Federal do Rio de Janeiro

RESUMO - A variância média extraída (VME) e a confiabilidade composta (CC) são indicadores associados à qualidade de uma medida. No entanto, é necessário compreender adequadamente a dinâmica dos cálculos da VME e da CC, bem como as suas relações com os conceitos de validade e precisão para evitar equívocos na interpretação dos seus resultados. No presente artigo, ilustramos, por meio de modelos unifatoriais simulados, como o número de itens e a homogeneidade das cargas fatoriais impactam os valores da VME e da CC. Assim, problematizamos a utilização de pontos de cortes fixos para esses indicadores. Além disso, apresentamos argumentos endossando a VME como uma medida de precisão, e não de validade convergente.

Palavras-chave: variância média extraída, confiabilidade composta, validade, precisão, análise fatorial confirmatória

Average Variance Extracted and Composite Reliability: Reliability Coefficients

ABSTRACT - The average variance extracted (AVE) and the composite reliability coefficients (CR) are related to the quality of a measure. Meanwhile, in order to avoid misconceptions, it is required to properly comprehend the equations of the AVE and CR, as well as their relation with the definition of validity and reliability. In this paper, we illustrate, using simulated one-factor models, how the number of items and the homogeneity of factor loadings might influence the AVE and CR results. In so doing, we show that the use of fixed AVE and CR cutoffs is problematic. Moreover, we present reasons endorsing the use of the AVE as a reliability coefficient, instead of a convergent validity index.

Keywords: Average variance extracted, composite reliability, validity, reliability, confirmatory factor analysis

Na área da Psicologia, o número de artigos científicos que busca compreender fenômenos psicológicos por meio de abordagens quantitativas é considerável. Entre esses, uma ampla parcela preocupa-se com a adaptação e validação de instrumentos psicométricos. Conforme literatura especializada (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014; Messick, 1989; Nunnally & Bernstein, 1994; Urbina, 2007), para serem considerados adequados, os escores estimados por meio de uma escala psicométrica devem apresentar diversas evidências de validade e, simultaneamente, indicadores adequados de precisão. Tais evidências devem ser testadas incessantemente, garantindo a aplicabilidade da escala em diversos contextos, diversas amostras e populações, bem como através do tempo (Sireci, 2007; Thompson, 2002).

Entre os conceitos de validade, a literatura clássica tem sugerido, ao menos, duas perspectivas dominantes. A primeira refere-se ao modelo tripartite, proposto em 1966 pelos Standards for Educational and Psychological Testing da *American Education Research Association* (AERA), *American Psychological Association* (APA) e *National Council on Measurement in Education* (NCME; AERA, APA, & NCME, 1966). Nessa estruturação, os conceitos de validade são distinguidos em validade de conteúdo, de construto e de critério (modelo CCC). Na validade de construto, encontra-se o conceito de validade convergente, que se refere à evidência de que os escores em um teste relacionam-se significativamente com os escores de um outro teste externo que mensura o mesmo atributo (Campbell & Fiske, 1959). Ainda que antigo, o modelo CCC tem sido costumeiramente utilizado na literatura (Urbina, 2007). Isto porque, a separação entre validade de conteúdo, construto e critério auxilia pesquisadores a compreenderem didaticamente as diferentes formas de validade.

Discussões posteriores à década de 1950 indicaram que a validade convergente pode ser dividida em validade de traço (*trait validity*) e validade nomológica (*nomological validity*). A primeira enfatiza as correlações entre medidas que supostamente devem avaliar o mesmo construto. Já a validade nomológica se refere à rede de relações que o construto estabelece com outros construtos correlatos, embora diferentes (Geisenger, 1992; Messick, 1980).

1 Apoio: Capes

2 Agradecimentos: Os autores agradecem aos pesquisadores Barbara Byrne, Cristiano M. Gomes, Dragos Iliescu, Igor G. Menezes, Jacob A. Laros, Nathan Thompson, Ronald Hambleton, e Stephen Sireci pelos valiosos comentários e discussões acerca dos conceitos de validade, precisão, VME e CC, que nos auxiliaram a consolidar a nossa linha argumentativa apresentada no presente artigo. Ademais, agradecemos ao Cristiano M. Gomes e ao Wagner de Lara. Machado pela revisão cuidadosa e pelos comentários construtivos sugeridos em versões anteriores deste manuscrito.

3 Endereço para correspondência: Universidade Salgado de Oliveira (UNIVERSO), Programa de Pós-Graduação em Psicologia, Rua Marechal Deodoro, 263 (Bloco A), Bairro Centro, Niterói, RJ. Brasil. CEP: 24.020-420. *E-mail*: valentini.felipe@gmail.com.

O segundo modelo dominante sobre os critérios de validade dos escores refere-se a uma atualização dos *Standards* de 1966, publicada em 1999 e revisada em 2014 (AERA, APA, & NCME, 2014). Desde a publicação de 1999, o conceito de validade passa a ser dividido em cinco fontes de evidências: (a) Evidências baseadas no conteúdo do teste; (b) Evidências baseadas nos processos de resposta; (c) Evidências baseadas na estrutura interna; (d) Evidências baseadas em relações com outras variáveis (validade convergente, discriminante, de critério e de generalização); e (e) Evidências baseadas nas consequências da testagem.

Um ponto importante a ser destacado nas diretrizes postuladas pelos Standards é que, a partir de 1999, a fidedignidade de um teste passa a ser compreendida como um critério de validade, enquanto que, nas versões anteriores desse *guideline*, a fidedignidade não se encontrava inserida no modelo de validade vigente (concebido, até então, pelo modelo CCC). Por outro lado, salienta-se que a definição do termo validade convergente se mantém estável desde a primeira versão dos *Standards*, em 1966, e se refere às associações que o escore em um teste apresenta com outras medidas externas. Ou seja, por meio da validade convergente busca-se avaliar até que ponto os escores dos sujeitos no instrumento X encontram-se relacionados com os escores dos mesmos sujeitos no instrumento Y, de modo que validade convergente, consagradamente, é compreendida como uma medida de relação r_{xy} (AERA, APA, & NCME, 1966, 1999, 2014; Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Messick, 1989; Urbina, 2007).

Embora haja um grande número de pesquisadores endossando essa perspectiva, é possível observar que o uso dos termos e a forma de se atestar validade convergente e fidedignidade não são consensuais na literatura. Em 1981, Fornell & Larcker (1981) publicaram um artigo, apresentando fórmulas alternativas de se calcular índices de fidedignidade e de validade convergente: a confiabilidade composta (CC) e o índice de variância média extraída (VME). A utilização desses indicadores é crescente na literatura científica e pode ser facilmente verificada em pesquisas empíricas recentes (e.g., Fock, Hui, Au, & Bond, 2013; Niclasen, Skovgaard, Andersen, Sømhovd, & Obel, 2013; Obasi, Brown, & Barrett, 2014). Entretanto, as discussões apresentadas por Fornell e Larcker, especificamente no que se refere à VME como indicador de validade convergente, não parecem ser condizentes com os conceitos postulados pela psicometria (AERA, APA, & NCME, 1999, 2014; Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Messick, 1989; Urbina, 2007).

Nesse contexto, o presente artigo tem como objetivo problematizar os conceitos de VME e CC, refutando a proposta de Fornell e Larcker (1981) de considerar a VME como indicador de validade convergente. Além disso, será demonstrado que os valores de VME e CC sofrem alterações em função do número de itens e da homogeneidade das cargas fatoriais, de modo que a utilização de pontos de corte fixos para esses indicadores podem limitar a interpretação dos resultados de um estudo empírico.

Definição e Equações da Confiabilidade Composta e da Variância Média Extraída

A CC e a VME são, primordialmente, indicadores que podem ser utilizados para avaliar a qualidade do modelo estrutural de um instrumento psicométrico (Hair, Black, Babin, Anderson, & Tatham, 2009; Fornell & Larcker, 1981). Para tanto, os cálculos da CC e VME são realizados com base nos parâmetros estimados por meio da Modelagem por Equações Estruturais (MEE). Ainda que os *softwares* mais utilizados para as análises MEE (AMOS, Mplus, EQS e Lisrel) não disponibilizem os resultados de CC e VME automaticamente em seus *outputs*, é possível obter esses indicadores por meio de cálculos simples. Para a CC, utiliza-se a seguinte equação (Fornell & Larcker, 1981):

$$CC = \frac{(\sum\lambda)^2}{(\sum\lambda)^2 + \sum\varepsilon} \quad \text{(equação 1)}$$

Na qual, CC é a confiabilidade composta; $\sum\lambda$ representa a soma das cargas fatoriais (ou coeficientes de regressão entre a variável latente e o item); e $\sum\varepsilon$ é a soma dos erros de mensuração (ou variância residual). Para modelos nos quais a variância e a média da variável latente foram fixadas (e todas as cargas fatoriais foram estimadas livremente), a equação pode ser montada tanto com os valores padronizados quanto com os valores não-padronizados, observado o paralelismo: caso se utilizem as cargas não-padronizadas, por exemplo, devem-se utilizar os erros de mensuração não-padronizados. As cargas fatoriais e os erros de mensuração não-padronizados podem ser obtidos diretamente no *output* do *software* de análise MEE. Por outro lado, para obter o erro de mensuração padronizado, é necessário utilizar o seguinte cálculo: um menos o quadrado da carga fatorial padronizada ($\varepsilon = 1 - \lambda^2$).

Ressalta-se que a equação da CC, apresentada por Fornell e Larcker (1981), pressupõe a independência dos erros. Ou seja, caso tenha-se estimado alguma covariância entre os resíduos (ou erros de mensuração), uma parte da variância residual (e consequentemente da variância total) será atribuída à covariância. Portanto, se o pesquisador estiver utilizando as cargas fatoriais não-padronizadas, é necessário inserir também a covariância residual no denominador da equação. Fornell e Larcker (1981) não discutiram esse aspecto no artigo original. No entanto, Raykov (2003), Rios e Wells (2014) apresentaram uma equação (bastante semelhante à fórmula de Fornell e Larcker) para estimar a precisão em modelos por equações estruturais, na qual a covariância residual é inserida no denominador da fração, como parte da variância total. Para modelos nos quais foi estimada alguma covariância residual diferente de 0, utiliza-se a seguinte equação:

$$CC_2 = \frac{(\sum\lambda)^2}{(\sum\lambda)^2 + \sum\varepsilon + 2\sum Cov} \quad \text{(equação 2)}$$

Na qual, CC_2 é a confiabilidade composta para modelos que incluem covariância residual; e $\sum Cov$ indica a soma das

covariâncias residuais; os demais termos já foram descritos na equação 1. Destacamos que a utilização de parâmetros padronizados ou não-padronizados, na equação 2, gera pequenas diferenças no valor da CC_2 (para modelos nos quais são estimadas covariâncias residuais). Sugere-se a utilização dos parâmetros padronizados, em conformidade com a apresentação da equação por Rios e Wells (2014). Ademais, o pesquisador deve sempre ponderar a pertinência da estimação de covariâncias residuais, pois tal estratégia reduz a precisão estimada pela CC. Ressalta-se ainda que, nos casos em que os erros são correlacionados, o tradicional coeficiente *alpha* pode sub ou superestimar a precisão dos escores (Raykov, 2003).

A CC tem sido apresentada como um indicador de precisão mais robusto, quando comparado ao coeficiente *alpha* (Cronbach, 1951). Isto porque, no cômputo da CC, as cargas ou pesos fatoriais dos itens são passíveis de variação, enquanto que, no coeficiente *alpha*, as cargas dos itens são fixadas para serem iguais, conforme postula o pressuposto da *tau*-equivalência (Raykov, 2001; Sijtsma, 2009). Nesse sentido, a CC tenderia a apresentar indicadores mais robustos de precisão por não estar atrelada a esse pressuposto (raramente observado empiricamente; Sijtsma, 2009).

Com o objetivo de investigar as reais diferenças encontradas nas estimativas de fidedignidade do coeficiente *alpha* e da CC, Peterson e Kim (2013) avaliaram o desempenho desses dois indicadores em pesquisas aplicadas. Por meio de uma meta-análise, os autores avaliaram 381 estudos (publicados em 327 artigos) que utilizaram, concomitantemente, o coeficiente alpha e a CC como indicadores de fidedignidade. Ao longo desses estudos, foram comparados 2524 pares de associações entre esses estimadores. Em média, a CC superou a magnitude do coeficiente alpha em 0,02 pontos ($\bar{X}$ *CC* = 0,86; $\bar{X}$ coeficiente *alpha* = 0,84), o que indica uma leve subestimação da precisão por meio do coeficiente *alpha* (*lower-bound to reliability*). A baixa diferença entre as estimativas desses dois índices de fidedignidade levou Peterson e Kim a concluírem que as discrepâncias encontradas apresentavam pouca ou nenhuma utilidade prática (para maiores informações, consultar Peterson & Kim, 2013).

Por outro lado, o CC pode ser bastante útil, do ponto de vista prático, para avaliação de modelos *bifactors*. Nesse contexto, é possível estimar a precisão dos escores dos fatores específicos, controlando a variância relacionada à dimensão geral. Caso a precisão seja baixa, a previsão de escores específicos pode estar embasada em informações pouco precisas. Salienta-se que esse procedimento não pode ser realizado utilizando o coeficiente alpha, o que justifica a relevância do cálculo de CC (para maiores informações, consultar Rios & Wells, 2014).

No artigo original, em 1981, Fornell e Larcker também discutem a VME. Para o cálculo, eles propõem a seguinte equação:

$$VME = \frac{\sum(\lambda^2)}{\sum(\lambda^2) + \sum\varepsilon} \quad \text{(equação 3)}$$

Na qual, VME é a Variância Média Extraída; λ^2 representa a carga fatorial elevada ao quadrado; consequentemente, $\sum(\lambda^2)$ indica a soma das cargas fatoriais elevadas ao quadrado; e $\sum\varepsilon$ é a soma dos erros de mensuração. Ressalta-se que, para o cálculo da VME, devem ser utilizadas as cargas fatoriais padronizadas. Na equação da VME, o erro de mensuração e o quadrado das cargas fatoriais são indicados na mesma unidade de medida, portanto, a VME representa a proporção média da variância dos itens explicada pela variável latente (ou traço/fator comum entre os itens).

A VME também pode ser compreendida simplesmente como a média das cargas fatoriais padronizadas ao quadrado. Considerando que, normalmente, $\lambda^2 + \varepsilon = 1$, se utilizadas as cargas fatoriais padronizadas, o denominador da equação 3 será um multiplicado pelo número de itens. Portanto, exclusivamente ao considerar as cargas fatoriais padronizadas, o denominador da equação 3 pode ser substituído pelo número de itens, e a referida equação pode ser simplificada:

$$VME = \frac{\sum(\lambda_p{}^2)}{k} \quad \text{(equação 4)}$$

Na qual, $\lambda_p{}^2$ representa o quadrado da carga fatorial padronizada; e k indica o número de itens. A VME pode ser interpretada como a quantidade média da variância dos itens explicada pela variável latente. Sabendo que a variância do item é explicada pela carga fatorial e pelo erro, a VME representa, portanto, a porcentagem média da variância dos itens livre de erro de mensuração (erro referente à ausência de consistência interna).

As equações de CC e VME (1 e 3, respectivamente) são semelhantes e ambas são calculadas a partir das cargas fatoriais e do erro (ou variância residual) apenas. A única diferença entre as duas equações reside no fato de que, para a CC, deve-se somar todas as cargas fatoriais antes de elevar a soma ao quadrado $[(\sum(\lambda)^2]$; e, para a VME, deve-se calcular o quadrado das cargas fatoriais antes de somá-las $[(\sum(\lambda)^2]$. Embora essa diferença nas equações não represente uma alteração substancial do ponto de vista teórico-psicométrico, os valores de CC e VME sofrem distintas interferências das cargas fatoriais e do número de itens do modelo.

Considerações Acerca do Estabelecimento de Pontos de Corte para Interpretação da CC e da VME

A discussão sobre a relação entre CC, VME, cargas fatoriais e número de itens é relevante para a ponderação de pontos de corte de CC e VME. Fornell e Larcker (1981) apresentaram o valor de VME igual ou superior a 0,50 como um indicador de ajuste adequado do modelo. Nesse caso (VME ≥ 0,50), a média das cargas fatoriais seria de, aproximadamente, 0,70. Para a CC, o ponto de corte não pode ser interpretado de maneira tão direta. Fornell e Larcker (1981) não discutem pontos de corte para esse indicador. Entretanto, outros autores recomendam o valor de 0,70 (Hair et al., 2009) ou ainda de 0,60 (Bagozzi & Yi, 1988) como

aceitável. De todo modo, um ponto de corte único e fixo, especialmente para a CC, não parece justificável devido à sua variabilidade em função do número de itens do instrumento e das cargas fatoriais.

Para melhor problematizar essa discussão, foram simuladas as cargas fatoriais de 64 modelos unidimensionais. As cargas fatoriais foram geradas por meio do software WinGen (Han, 2007; Han & Hambleton, 2007). No conjunto de dados simulados, 30 modelos foram estimados com diferentes números de itens (5, 10 e 30, respectivamente, sendo 10 bancos simulados para cada modelo). Nesses 30 modelos, a variância das cargas fatoriais foi fixada a 0 (modelos completamente homogêneos). Por meio de tais modelos, buscou-se ilustrar as diferenças entre VME e CC para modelos com variados números de itens (controlando o efeito da heterogeneidade das cargas fatoriais). Os 34 modelos remanescentes foram simulados com diferentes variações da heterogeneidade das cargas fatoriais (desvios-padrão iguais a 0,00/ 0,10/ 0,20/ 0,30/ 0,40) e número de itens fixo em 10. Por meio desses 34 modelos, buscou-se ilustrar o impacto da heterogeneidade das cargas fatoriais na estimação da VME e CC.

É importante salientar que, em alguns casos, as cargas fatoriais foram ajustadas manualmente para maior controle da variância, considerando que o *software* nem sempre forneceu valores exatamente iguais aos solicitados. Por exemplo, se, ao solicitar ao *software* um conjunto de cargas fatoriais com desvio-padrão (*DP*) igual a 0,10, o *software* simulava um conjunto com *DP* igual a 0,11, algumas cargas fatorais, nesses casos, foram ajustadas manualmente para que o *DP* fosse exatamente igual a 0,10.

A Figura 1 apresenta os valores de CC e VME para diferentes médias de cargas fatoriais padronizadas em função do número de itens. A VME não sofre influência do número de itens do modelo, uma vez que o número de itens é ponderado no cálculo, como único valor do denominador da equação (ver equação 4). Por outro lado, a CC é moderada pelo número de itens, além das cargas fatoriais. Considerando que as cargas fatoriais são somadas antes de serem elevadas ao quadrado (ver equação 1), essa soma será alterada em função do número de itens; consequentemente, o aumento da soma das cargas fatoriais (provocado pelo aumento do número de itens) resultará no aumento do valor de CC.

Ademais, a Figura 1 evidencia que as relações entre CC, VME e cargas fatoriais não são lineares, embora positivamente relacionadas (correlações entre CC e VME: para cinco itens, r = 0,92; para dez itens, r = 0,85; para 30 itens, r = 0,73). Considerando que a base do cálculo da VME são as cargas fatoriais ao quadrado, as maiores alterações dos valores de VME ocorrem entre modelos com cargas fatoriais altas. Por exemplo, em dois modelos com médias de cargas fatoriais padronizadas iguais a 0,2 e 0,3, os valores de VME são, respectivamente, 0,04 e 0,09 (diferença de 0,05). Já para modelos com médias de cargas fatoriais iguais a 0,80 e 0,90, os valores de VME são, respectivamente, 0,64 e 0,81 (diferença de 0,17). Nesse sentido, a VME é mais sensível aos modelos com cargas fatoriais altas e tende a rejeitar, indiscriminadamente, os modelos com cargas fatoriais baixas. Por outro lado, a CC é mais sensível às variações de modelos com cargas fatoriais baixas. Por exemplo, para os modelos com médias de cargas fatoriais iguais a 0,10 e 0,20 (ambos com dez itens), a diferença das CC dos modelos é de 0,20 (CC de 0,10 e 0,30, respectivamente).

Observa-se na Figura 1, ainda, que o valor de CC sofre menos influência das médias das cargas fatoriais com o acréscimo de itens ao modelo. Por exemplo, para um modelo de cinco itens, o valor de CC de 0,70 é obtido se a média das cargas fatoriais padronizadas for de aproximadamente 0,60. Contudo, se aumentarmos o modelo para 30 itens, o valor de CC de 0,70 é alcançado com média de cargas fatoriais de aproximadamente 0,30; uma média considerada baixa para um modelo de equações estruturais (mesmo que o valor de CC seja de 0,70). Assim, para fatores com muitos itens (acima de 10), sugere-se considerar a possibilidade de adotar um ponto

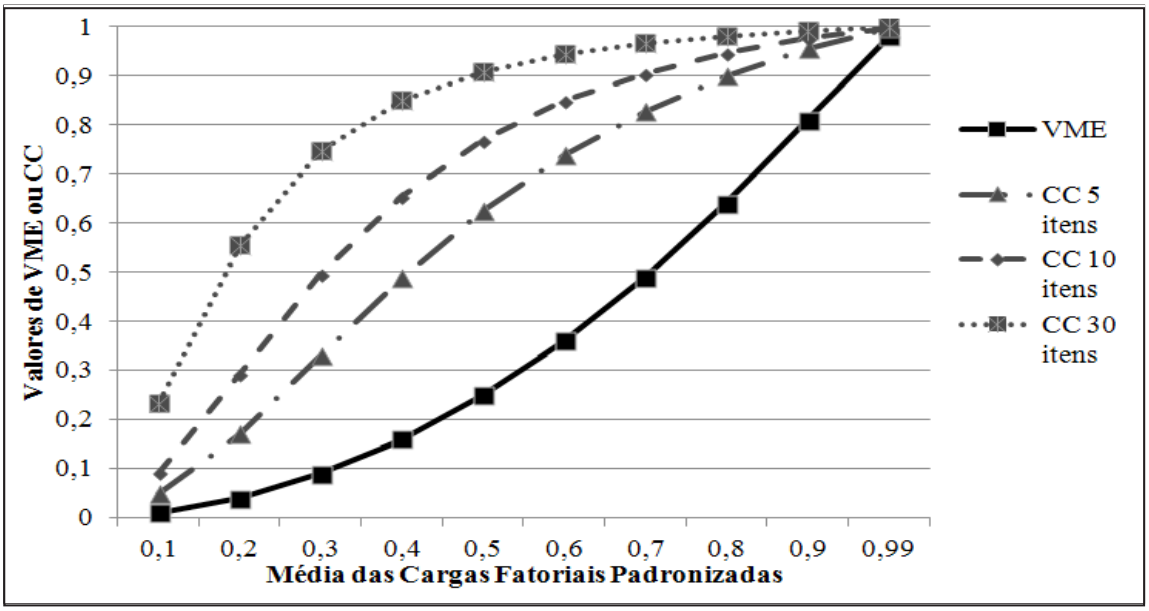

Figura 1. Relações entre os valores de CC e de VME e a média das cargas fatoriais em função do número de itens.

de corte mais conservador. Sabe-se que todos os indicadores de precisão tendem a aumentar com o acréscimo de itens no modelo. Contudo, os pesquisadores precisam estar atentos à relatividade dos pontos de corte, pois é possível que estejam validando um modelo mesmo com cargas fatoriais baixas (se tiverem um elevado número de itens).

Outro aspecto importante que deve ser considerado ao discutir os pontos de corte de VME e CC diz respeito à homogeneidade das cargas fatoriais do modelo. Considerando que a homogeneidade pode ser avaliada pelo desvio-padrão das cargas fatoriais, avaliou-se a influência dos desvios-padrões na estimação nos valores de CC e VME (mantendo constante o número de 10 itens). A Figura 2 apresenta esses resultados. Os valores de CC sofrem pequena influência da heterogeneidade das cargas fatoriais (*DP* altos): normalmente a diferença é inferior a 0,05 entre um modelo com cargas fatoriais perfeitamente homogêneas (*DP* = 0) e outro modelo com cargas fatoriais heterogêneas (*DP* = 0,40). Por exemplo, para modelos com dez itens e média das cargas fatoriais padronizadas de 0,70, a CC é igual a 0,91 para o modelo com cargas perfeitamente homogêneas; para um modelo cujas cargas apresentam elevada heterogeneidade, o valor de CC é igual a 0,93. Em outras palavras, para o exemplo, a diferença de CC entre um modelo com cargas homogêneas e outro com cargas heterogêneas é de apenas 0,02.

Por outro lado, a dinâmica da VME em relação à homogeneidade é bastante diferente: a VME tende a inflar com o aumento da heterogeneidade das cargas fatoriais. Por exemplo, para modelos com 10 itens e média de cargas fatoriais padronizadas de 0,60, a VME é igual 0,36, caso as cargas fatoriais sejam completamente homogêneas (*DP* = 0); já para um modelo com média de cargas fatoriais padronizadas também de 0,60, mas com elevada heterogeneidade (*DP* = 0,40), a VME passa a ser de 0,52. Nesse exemplo, mantendo constante o número de itens e a média de cargas fatoriais, a VME variou 0,15 entre o modelo homogêneo e o heterogêneo. Ademais, percebe-se na Figura 2 que o valor de VME pode ser igual (ou superior) a 0,50 mesmo a partir de cargas fatoriais média abaixo de 0,70 (para modelos com cargas muito heterogêneas, *DP* = 0,40). Portanto, recomendamos cautela na interpretação dos valores de VME para modelos cujas cargas fatoriais apresentem grande variabilidade, pois, nesse caso, a VME pode superestimar o ajuste do modelo. Além disso, sugerimos que os pesquisadores avaliem o desvio-padrão das cargas fatoriais e, caso esse valor seja alto, considerem a possibilidade de adotar um ponto de corte de VME mais conservador. Por exemplo, para um modelo hipotético, cujas cargas fatoriais apresentam *DP* igual a 0,30 (modelo com cargas heterogêneas), a VME deve ser igual a 0,59, no mínimo, para que a média das cargas fatoriais seja superior a 0,70 (conforme Figura 2). Portanto, para esse exemplo, o ponto de corte fixo de VME maior ou igual a 0,50 não é justificável.

Considerando os problemas apresentados sobre a utilização de pontes de corte fixos para avaliação da CC e da VME, sugere-se ponderar estes valores considerando o número de itens e a homogeneidade das cargas fatoriais do modelo. Os gráficos apresentados nas figuras 1 e 2 podem auxiliar os pesquisadores a definir os pontos de corte mais adequados para seu modelo estrutural.

Confiabilidade Composta e Variância Média Extraída: Indicadores de Precisão

A CC e a VME têm sido apresentadas e discutidas na literatura científica sob diversos pontos de vista, principalmente no que se refere à utilização dos termos "validade" e "precisão". Ainda que Fornell e Larcker (1981) tenham utilizado o termo "validade convergente" para a VME, julgamos que esse uso do termo não é coerente com a psicometria (e.g., AERA, APA, & NCME, 2014; Messick,

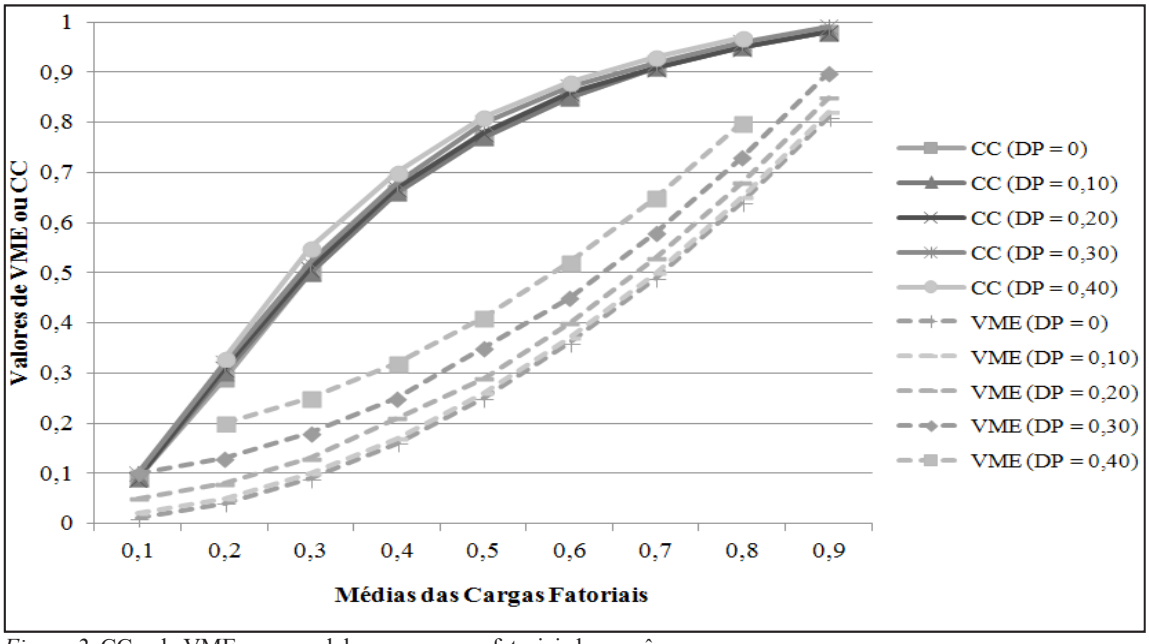

Figura 2. CC e de VME para modelos com cargas fatoriais homogêneas e heterogêneas.

1989; Nunnally & Bernstein, 1994; Urbina, 2007), em função de dois argumentos principais.

O primeiro argumento diz respeito às equações 1 e 3 disponibilizadas neste artigo. Conforme já discutido, ainda que as equações produzam indicadores distintos, as diferenças nas equações são pequenas e não há justificativa teórica para tratar a CC como precisão e a VME como validade convergente, ao menos do ponto de vista da psicometria (AERA, APA, & NCME, 2014; Messick, 1989; Nunnally & Bernstein, 1994).

O segundo argumento diz respeito ao uso dos termos validade convergente e "precisão", utilizados por Fornell e Larcker (1981), e ao entendimento que a psicometria tem sobre esses indicadores (AERA, APA, & NCME, 2014). Em diversas partes do artigo original, Fornell e Larcker (1981) utilizaram os termos "validade convergente" e "precisão" indistintamente, inclusive como sinônimos. Por vezes, os autores discutem a "validade convergente" contrapondo-a ao erro de mensuração (ou seja, a "validade convergente" representaria a parte dos itens que é livre do erro de mensuração e capturada pela variável latente). Entretanto, na psicometria, "erro de mensuração" está diretamente associado à precisão dos escores (AERA, APA, & NCME, 1999; Cronbach & Meehl, 1955; Nunnally & Bernstein, 1994; Sijtsma, 2009). Assim, o uso do termo "validade convergente", no artigo de Fornell e Larcker (1981), faz explicita alusão ao que a psicometria entende por precisão dos escores, ou seja, fidedignidade ou confiabilidade.

Entretanto, embora relacionados e complementares, precisão e validade convergente são aspectos distintos. Conforme mencionado no início do presente estudo, validade convergente é definida pela relação entre uma medida "X" e outra medida externa "Y", pressupondo que a medida "Y" avalia, de maneira válida, o mesmo construto da medida "X" (*trait validity*) ou algum construto correlato (*nomological validity*) (Geisenger, 1992). Corroborando essa perspectiva, Nunnaly e Bernstein (1994), por exemplo, afirmam que a validade convergente "demonstra que dois métodos independentes para inferir um atributo conduzem a conclusões semelhantes" (pp. 92).

A modelagem multitraço-multimétodo (MTMM), proposta por Campbell e Fiske (1959), também advoga em favor da nossa linha argumentativa, de que validade convergente necessita de uma medida externa. Campbell e Fiske (1959) propuseram um modelo clássico para analisar, simultaneamente, a validade convergente e discriminante, bem como a precisão dos escores. Nesse artigo, os autores indicaram que a validade convergente é uma questão "heterométodo" (mais de um método) e reservaram o conceito de precisão às relações "monotraço-monométodo" (único traço e único método). A modelagem MTMM foi adaptada ao contexto da modelagem por equações estruturais na década de 1980 (Wildaman, 1985). Essa modelagem é conduzida por meio de comparações de diferentes modelos, os quais apresentam, no mínimo, dois traços e dois métodos distintos (Wildaman, 1985). Considerando que a VME é calculada para um traço latente apenas, não parece razoável argumentar a existência de uma "medida externa". Ademais, Fornell e Larcker (1981) pareciam mais preocupados em defender a posição de que, se os itens "convergem" entre si (e, por consequência, a VME é alta), a estrutura do modelo é adequada (por isso "validade convergente do construto")[1]. Nesse sentido, a VME encontra-se fortemente associada à noção clássica de correlação medida-a-medida (representada por rxx), na qual as relações entre "partes" (ou itens) de um mesmo instrumento indicam o quão estável é a medida. Considerando que as diferentes "partes" deveriam avaliar o mesmo construto, a ausência de correlação (r_{xx}) é interpretada como instabilidade (ou imprecisão).

Ao observar as equações apresentadas por Fornell e Larcker (1981), pode-se perceber que o cálculo da VME é semelhante à proposta da "Teoria do Escore Verdadeiro", da Teoria Clássica dos Testes (TCT; Ferguson, 1981; Guilford, 1954; Primi, 2012). Nessa perspectiva, postula-se que a precisão dos escores é indicada pela proporção do escore total que é devida ao "escore verdadeiro" (portanto, Precisão = V / T, no qual V representa a variância verdadeira e T, a variância total). No cálculo da VME, o numerador da equação refere-se à soma das cargas fatoriais ao quadrado, o que representa a variância compartilhada entre o traço e os itens, ou seja, o quanto os itens avaliam o traço. Essa soma pode ser igualmente interpretada como a variância verdadeira dos escores. O denominador da equação da VME, por sua vez, refere-se à soma das cargas fatoriais ao quadrado mais o erro, o que pode ser interpretado como a variância total. Considerando que a soma das cargas fatoriais ao quadrado é o numerador da equação da VME e a variância total é o denominador, a equação da VME é idêntica à noção de precisão da Teoria do Escore Verdadeiro: variância verdadeira dividida por variância total. Essa é uma justificativa clara para compreender a VME como um indicador de precisão, e não de validade convergente.

Conclusão

Ao longo deste manuscrito, apresentamos uma série de argumentos, com base em modelos de equações estruturais simulados, que advogam em favor da relativização do uso de pontos de corte fixos para a VME e para a CC. Os valores de ambos os indicadores sofrem alterações em função do número de itens e da homogeneidade das cargas fatoriais, de modo que a utilização de pontos de corte fixos pode limitar a interpretação dos resultados de um estudo empírico. Embora não seja possível, a partir deste estudo, estabelecer novos pontos de corte para a CC e a VME, as informações apresentadas e sintetizadas nas Figuras 1 e 2, poderão servir como base para que os pesquisadores avaliem criticamente os seus modelos, de acordo com suas características específicas, e estabeleçam pontos de corte mais adequados para essas características.

Além disso, esse estudo buscou problematizar a utilização da VME como indicador de validade convergente. Na nossa

1 Ressalta-se que a VME também pode ser utilizada para avaliar a validade discriminante entre duas variáveis latentes em uma modelagem por equações estruturais. Nesse sentido, comparam-se os valores de VME com a correlação ao quadrado das dimensões. Caso as VMEs apresentem valores superiores à correlação ao quadrado, isto indica que os construtos são minimamente diferentes (validade discriminante). Para maiores informações, ver Fornell e Larcker (1981) e Farrell (2010).

compreensão, embora com métricas distintas, tanto a VME quanta a CC apresentam evidências de precisão dos escores. É importante salientar, porém, que atualmente a precisão dos escores é considerada, de fato, uma evidência de validade. De todo modo, a última versão dos Standards (AERA, APA, NCME, 1999, 2014) apresenta o conceito de precisão como uma evidência de validade da estrutura interna, mas reserva o conceito de validade convergente para a relação entre medidas diferentes.

Referências

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1966). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*(1), 74–94. doi:10.1007/BF02723327

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. doi:10.1037/h0046016

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. doi:10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. doi:10.1037/h0040957

Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research, 63*(3), 324-327. doi:10.1016/j.jbusres.2009.05.003

Ferguson, G. A. (1981). *Statistical Analysis in Psychology and Education*. New York, NY: McGraww-Hill

Fock, H., Hui, M. K., Au, K., & Bond, M. H. (2013). Moderation effects of power distance on the relationship between types of empowerment and employee satisfaction. *Journal of Cross-Cultural Psychology, 44*(2), 281–298. doi:10.1177/0022022112443415

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equations models with unobservable variables and measurement error. *Journal of Marketing, 18*(1), 39-50. doi:10.2307/3151312

Geisenger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist, 27*(2), 197-222. doi: 10.1207/s15326985ep2702_5

Guilford, J. P. (1954). *Psychometric Methods*. New York, NY: McGraw-Hill

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados* (6th. ed.). Bookman: Porto Alegre.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459. doi:10.1177/0146621607299271

Han, K. T., & Hambleton, R. K. (2007). *User's Manual: WinGen* (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts, School of Education.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027. doi:10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Niclasen, J., Skovgaard, A. M., Andersen, A.-M. N., Sømhovd, M. J., & Obel, C. (2013). A confirmatory approach to examining the factor structure of the Strengths and Difficulties Questionnaire (SDQ): A large scale cohort study. *Journal of Abnormal Child Psychology, 41*(3), 355–365. http://doi.org/10.1007/s10802-012-9683-y

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd. ed.). New York: McGraw-Hill.

Obasi, C. N., Brown, R. L., & Barrett, B. P. (2014). Item reduction of the Wisconsin Upper Respiratory Symptom Survey (WURSS-21) leads to the WURSS-11. *Quality of Life Research, 23*(4), 1293–1298. http://doi.org/10.1007/s11136-013-0561-z

Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. J*ournal of Applied Psychology, 98*(1), 194–198. doi:10.1037/a0030767

Primi, R. (2012). Psicometria: Fundamentos matemáticos da Teoria Clássica dos Testes. *Avaliação Psicológica, 11*(2), 297-307. Retirado de: http://pepsic.bvsalud.org/pdf/avp/v11n2/v11n2a15.pdf

Raykov, T. (2001). Bias of coefficient for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*(1), 69–76. doi:10.1177/01466216010251005

Raykov, T. (2003). *Scale reliability evaluation with LISREL 8.50*. Retrieved from www.ssicentral.com/lisrel/techdocs/reliabil.pdf

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26(*1), 108-116. doi:10.7334/psicothema2013.260

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach"s alpha. *Psychometrika, 74*(1), 107-120. doi:10.1007/s11336-008-9101-0

Sireci, S. G. (2007). On validity and test validation. *Educational Researcher, 36*(8), 477-481. doi:10.3102/0013189X07311609

Thompson, B. (2002). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage publications.

Urbina, S. (2007). *Fundamentos da testagem psicológica*. Porto Alegre: Artmed.

Wildaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*(1), 1-26. doi:10.1177/014662168500900101

Recebido em 11.08.2014
Primeira decisão editorial em 05.04.2015
Versão final em 04.05.2015
Aceito em 20.05.2015 ■