Editorial

DOI: 10.15689/ap.2017.1601.ed

Em nome da Revista **Avaliação Psicológica** e do Instituto Brasileiro de Avaliação Psicológica (IBAP), inicio esta edição agradecendo à ilustre Profª. Dra. Acácia Aparecida Angeli dos Santos pelos seus mais de seis anos à frente da editoração deste periódico. Os esforços da Profª. Acácia e de sua equipe, somados ao trabalho dos colegas que a precederam na função, foram decisivos para que a Revista se tornasse um dos principais veículos de divulgação científica latino-americano na área da avaliação psicológica. É por isso que, apesar da mudança de editor chefe, a política editorial, o escopo e a missão da revista permanecem inalterados. Inspirados pelo exemplar trabalho da Profª. Acácia, esperamos poder continuar no caminho por ela trilhado. Nossa equipe de trabalho agora conta com a ajuda do editor Associado Prof. Dr. Felipe Valentini, da Editora Júnior Profa. Dra. Francine Nathalie Ferraresi Rodrigues Queluz, e da assistente editorial Adriana Ferraz Satico.

O presente editorial também aborda, superficial e brevemente, a problemática do teste empírico de modelos teóricos em Psicologia. Após uma breve introdução, são discutidos dois pontos específicos relacionados com a temática.

A testabilidade (isto é, "falseabilidade") de uma teoria é um dos critérios que separam a ciência da não ciência (Popper, 1959). Em geral (embora nem sempre), testar as hipóteses e as implicações de uma teoria requer que ela seja formulada matematicamente, caso em que recebe o nome de "modelo" (Rodgers, 2010). No caso da Psicologia, modelos de equações estruturais são a abordagem padrão, uma vez que permitem endereçar, simultaneamente, diversas predições e hipóteses de relacionamento entre variáveis observadas ou latentes.

Naturalmente, modelos não são um *buffet*, em que é possível escolher e colocar no prato o que mais agrada ao paladar da pesquisadora ou do pesquisador. Modelos devem ser testados empiricamente contra dados reais, a fim de determinar se conseguem se sair bem como uma tentativa de desvendar a verdadeira estrutura causal oculta que produz os fenômenos aparentes. Entretanto, está longe de haver um consenso sobre qual a maneira "correta" de avaliar o ajuste de modelos de equações estruturais aos dados.

Uma das maiores controvérsias na área diz respeito ao uso dos índices de ajuste aproximado em substituição ao teste do qui-quadrado (*Confirmatory Fit Index, Tucker-Lewis Index, Root Mean Square Error of Approximation*, entre outros). Grosseiramente falando, em modelos de equações estruturais, a hipótese nula do teste do qui-quadrado é a de que a matriz de variâncias-covariâncias implicada pelo modelo reproduz perfeitamente a matriz empírica dos dados. Supostamente, o teste do qui-quadrado tende a identificar como significativas ($p < 0,05$) mesmo diferenças pequenas entre essas duas matrizes quando o tamanho amostral é grande. Isso motivou a criação de índices de ajuste, coeficientes alternativos que são mais permissivos, e podem revelar o ajuste "aproximado" do modelo aos dados. O embate qui-quadrado *versus* índices de ajuste chegou a produzir um número especial na *Personality and Individual Differences*, em 2007 (ver Barrett, 2007; Bentler, 2007; Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007; McIntosh, 2007), a partir de discussões iniciadas na SEMnet.

O presente editorial ressalta dois pontos importantes no debate. O primeiro deles é que as controvérsias na área não estão resolvidas, e que índices de ajuste não "venceram" a batalha contra o qui-quadrado. Há demonstrações de que o qui-quadrado só aumenta relativamente aos graus de liberdade quando o modelo em questão é falso (Hayduk, 2014); ou seja, se o modelo hipotetizado é equivalente ao modelo verdadeiro, o teste do qui-quadrado tende a ser não significativo, mesmo com um tamanho amostral muito grande.

[1]Sintaxe R disponível mediante requisição pelo e-mail hauck.nf@gmail.com.

Essa premissa pode ser ilustrada facilmente por meio de uma simulação de dados. Para exemplificar o argumento, o pacote *simsem* foi empregado para gerar um banco de dados com 5.000 casos, tendo como modelo verdadeiro um modelo unidimensional com dez indicadores dicotômicos[1]. Ao especificar corretamente o modelo descrito acima em uma análise confirmatória com os dados gerados, o teste do qui-quadrado não resultou significativo tanto ao usar um estimador robusto (quadrados mínimos ponderados) — $\chi(35)=41{,}61$, $p=0{,}205$, CFI=0,996, TLI=0,995, RMSEA=0,014 — quanto ao empregar o tradicional *Maximum Likelihood*, menos indicado para itens dicotômicos — $\chi(35)=42{,}74$, $p=0{,}173$, CFI=0,992, TLI=0,990, RMSEA=0,015. Ainda, o valor *p* não se tornou significativo mesmo ao aumentar o tamanho amostral em dez vezes (50 mil casos). Em outras palavras, não é verdade que o valor *p* do teste do qui-quadrado será significativo mesmo para o modelo verdadeiro em situações de grande tamanho amostral.

O segundo ponto é que o debate teste do qui-quadrado *versus* índices de ajuste esconde algo muito mais interessante, que é a "especificação teórica do modelo". A estrutura causal no mundo real dificilmente é tão simples quanto um fator, sendo a única explicação para dez comportamentos diferentes em um grupo de 5.000 pessoas com histórias distintas (o exemplo da simulação acima). Muitas outras influências podem existir no nível dos respondentes, dos itens e do contexto. Apenas para ilustrar, no nível dos respondentes, podem desempenhar um papel o sexo ou o gênero (Millsap, 2011), disposições pessoais para realçar qualidades e esconder defeitos (Paulhus & John, 1998) e também uma série de estilos de resposta (Van Vaerenbergh & Thomas, 2013). No nível do item, podem ser variáveis de impacto o tamanho do enunciado do item (Hamby & Ickes, 2015), o uso de adjetivos com flexão de gênero (Vainapel, Shamir, Tenenbaum, & Gilam, 2015) e a presença de itens negativamente relacionados com o traço latente (Schriesheim & Eisenbach, 1995). No nível do contexto da avaliação, pode ocorrer a influência de incentivos para a simulação de respostas (Ziegler & Buehner, 2009) e de uma série de outras variáveis, incluindo até mesmo condições meteorológicas do dia da aplicação do instrumento (Rammstedt, Mutz, & Farmer, 2015). Apesar de numerosos, esses exemplos não esgotam a lista de influências que participam na variância sistemática de um conjunto de dados psicológicos.

A questão é que modelos que não consideram a possibilidade de algumas dessas influências podem não estar corretamente especificados. Isso, por sua vez, pode produzir estimativas paramétricas enviesadas, independentemente do tamanho amostral empregado (Antonakis, Bendahan, Jacquart, & Lalive, 2010). Em outras palavras, o programa estatístico não vai resolver um problema de entendimento teórico da relação entre as variáveis, e os índices de ajuste aproximado pouco podem fazer para remediar a situação. Apesar de se chamarem técnicas "confirmatórias", o mero uso das equações estruturais não garante qualquer prerrogativa ao modelo testado. Como de praxe nas ciências, nada está sendo confirmado ou provado (Popper, 1959).

Concluindo, o teste do qui-quadrado não deve ser negligenciado, tampouco reificado. Precisamos é de modelos corretamente especificados: são esses que irão resistir a todos os testes empíricos, não importam quais sejam os critérios utilizados.

Nelson Hauck
Editor
Universidade São Francisco

Referências

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: a review and recommendations. *The Leadership Quarterly*, *21*(6), 1086-1120. http://doi.org/10.1016/j.leaqua.2010.10.010

Barrett, P. (2007). Structural equation modelling: adjudging model fit. *Personality and Individual Differences*, *42*(5), 815-824. http://doi.org/10.1016/j.paid.2006.09.018

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, *42*(5), 825-829. http://doi.org/10.1016/j.paid.2006.09.024

Hamby, T., & Ickes, W. (2015). Do the readability and average item length of personality scales affect their reliability? *Journal of Individual Differences*, *36*(1), 54-63. http://doi.org/10.1027/1614-0001/a000154

Hayduk, L. A. (2014). Shame for disrespecting evidence: the personal consequences of insufficient respect for structural equation model testing. *BMC Medical Research Methodology*, *14*(1), 124. http://doi.org/10.1186/1471-2288-14-124

Hayduk, L. A., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three – Testing the theory in structural equation models! *Personality and Individual Differences*, *42*(5), 841-850. http://doi.org/10.1016/j.paid.2006.10.001

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: a commentary and elaboration on Barrett (2007). *Personality and Individual Differences*, *42*(5), 859-867. http://doi.org/10.1016/j.paid.2006.09.020

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge, Taylor & Francis Group.

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: the interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*(6), 1025-1060. http://doi.org/10.1111/1467-6494.00041

Popper, K. R. (1959). *The logic of scientific discovery*. London, England: Routledge.

Rammstedt, B., Mutz, M., & Farmer, R. F. (2015). The answer is blowing in the wind: weather effects on personality ratings. *European Journal of Psychological Assessment*, *31*(4), 287-293. http://doi.org/10.1027/1015-5759/a000236

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *The American Psychologist*, *65*(1), 1-12. http://doi.org/10.1037/a0018326

Schriesheim, C. A., & Eisenbach, R. J. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management*, *21*(6), 1177-1193. http://doi.org/10.1177/014920639502100609

Vainapel, S., Shamir, O. Y., Tenenbaum, Y., & Gilam, G. (2015). The dark side of gendered language: the masculine-generic form as a cause for self-report bias. *Psychological Assessment*, *27*(4),1513-1519. http://doi.org/10.1037/pas0000156

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: a literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195-217. http://doi.org/10.1093/ijpor/eds021

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement*, *69*(4), 548-565. http://doi.org/10.1177/0013164408324469