Avaliação Psicológica
ISSN 2175-3431 (versão on-line)

Editorial

A análise de componentes principais é útil para selecionar bons itens quando a dimensionalidade dos dados é desconhecida?

Nelson Hauck, *Felipe Valentini*
Universidade São Francisco, Campinas-SP, Brasil

Um lugar-comum psicométrico é que, se um instrumento é novo ou dele pouca informação empírica se dispõe, então o primeiro passo deve ser conduzir uma análise exploratória. De fato, essa é uma etapa essencial e, para tanto, existem diferentes técnicas de redução de dados, como a análise fatorial exploratória e a análise de componentes principais (Black et al., 2012; Tabachnick & Fidell, 2007). Ao serem testados os itens de uma nova escala pela primeira vez, o desafio é duplo: o pesquisador quer, ao mesmo tempo, determinar a dimensionalidade dos seus dados e escolher os melhores indicadores. Este editorial enfatiza os perigos de conduzir a seleção de bons itens utilizando a análise de componentes principais quando a dimensionalidade dos dados é desconhecida.

A análise fatorial exploratória e a análise de componentes principais são métodos de redução de dados com propósitos distintos. Não é o objetivo deste editorial explorar essas diferenças em profundidade, dada a existência de literatura abundante sobre o assunto (e.g., Damásio, 2012; Gorsuch, 1990; Markus & Borsboom, 2013). Ainda assim, um contraste merece ser enfatizado. Enquanto a análise fatorial implica um modelo populacional que busca reproduzir a matriz empírica de variâncias e covariâncias dos itens (ou seja, há ênfase na *variância comum*), os componentes principais buscam apenas representar a *variância total* em um conjunto menor de variáveis (Yanai & Ichikawa, 2006). Disso decorre que, mantendo número de dimensões, itens e rotação constantes, a diferença entre os resultados dessas técnicas será tão maior quanto mais imperfeitos forem os indicadores (i.e., quanto mais erro de medida apresentarem). Por exemplo, se a carga verdadeira do item *j* no fator η for 0,30 ($0{,}30^2=9\%$ de variância compartilhada com o fator), então existirão 81% de variância não explicados por esse fator. Se essa variância remanescente for puramente erro, será considerada como tal na análise fatorial, mas poderá ser incorporada às cargas da solução de componentes principais. Gorsuch (1990) ilustrou como cargas componenciais espúrias podem surgir de uma situação como essa.

Neste breve estudo, ilustramos os problemas da seleção de itens com base em uma solução componencial quando a dimensionalidade dos dados é desconhecida. De maneira geral, exagerar o número de dimensões, isto é, cometer *overfactoring*, acarretará a manutenção de um modelo incorreto, com estimativas paramétricas enviesadas (Brown, 2015). Ainda assim, esse problema é potencialmente maior na análise de componentes principais, pois seu objetivo é representar a variância total dos dados com um mínimo de componentes ou índices. Se o pesquisador está explorando a qualidade de um novo conjunto de itens, a solução de componentes poderá levar à manutenção de itens pouco discriminativos.

Utilizando o pacote lavaan (Rosseel, 2012), foram simuladas as respostas de 500 indivíduos a 12 itens avaliativos de um fator latente. Enquanto os nove primeiros itens foram especificados para serem excelentes indicadores do fator, com carga=0,80, os três últimos foram especificados para serem pouco discriminativos, com carga de apenas 0,15. Por simplicidade de exposição, todos os itens foram criados para serem contínuos, com média 0 e desvio-padrão 1. Os dados simulados foram então analisados pela estimação *maximum likelihood* e pelos componentes principais com o pacote psych (Revelle & Revelle, 2015). A qualidade de soluções fatoriais e componenciais de um e dois fatores foi comparada. A ideia é que, se o pesquisador estivesse construindo uma escala, seria desejável manter os nove primeiros itens, mas excluir os três últimos.

A Tabela 1 apresenta os resultados das análises. Como se pode observar, tanto a análise fatorial *maximum likelihood* quanto a análise de componentes principais fizeram um bom trabalho em identificar os nove bons itens dentre o conjunto de dados quando apenas uma dimensão foi extraída. Em ambos os casos, o viés (diferença para com o valor do modelo verdadeiro) foi pequeno, especialmente para os bons itens. As cargas da solução *maximum likelihood* tiveram vieses para mais ou para menos, enquanto as cargas da solução componencial foram sistematicamente mais elevadas em 0,00–0,05. Contudo, uma diferença radical ocorreu entre as soluções de dois fatores. Verifica-se que a estimação *maximum likelihood* continuou identificando os nove primeiros itens como bons indicadores do fator F1, e o fator F2

como espúrio, sem uma associação linear ≥0,30 com quaisquer dos itens testados. Em contraste, a solução de componentes principais apontou os três últimos itens como bons indicadores de um segundo componente. Como o modelo verdadeiro continha apenas um fator, trata-se de uma dimensão espúria, mas que acabou sendo representada com três itens com carga >|0,45|. Com isso concluímos que, em uma situação real: 1. ao usar análise fatorial exploratória, o pesquisador faria a mesma e correta seleção de bons itens, mesmo que tivesse extraído um fator a mais dos seus dados; e 2. ao empregar componentes principais, a seleção dos itens seria influenciada pelo número de fatores extraídos, de modo que, errando a dimensionalidade dos dados, a seleção de itens espúrios provavelmente ocorreria.

Tabela 1
Modelo Verdadeiro e Parâmetros Estimados via Maximum Likelihood Versus Componentes Principais

	Modelo verdadeiro	ML–1f	ML–2f		CP–1c	CP–2c	
	F1	F1	F1	F2	F1	F1	F2
V1	0,80	0,80	0,80	–0,06	0,82	0,82	0,04
V2	0,80	0,78	0,78	0,12	0,81	0,81	–0,03
V3	0,80	0,80	0,80	0,08	0,83	0,83	–0,04
V4	0,80	0,81	0,82	–0,17	0,83	0,83	0,03
V5	0,80	0,82	0,81	0,15	0,84	0,84	–0,08
V6	0,80	0,79	0,80	–0,06	0,82	0,82	0,00
V7	0,80	0,80	0,80	–0,02	0,83	0,83	0,01
V8	0,80	0,80	0,80	0,05	0,82	0,82	0,02
V9	0,80	0,82	0,83	–0,07	0,84	0,84	0,04
V10	0,15	0,18	0,17	0,22	0,20	0,20	–0,67
V11	0,15	0,20	0,21	–0,12	0,23	0,23	0,59
V12	0,15	0,13	0,13	0,00	0,15	0,15	0,45
r			0,03			–0,01	

Nota. ML–1f=modelo fatorial com um fator e estimação *maximum likelihood*, ML–2f=modelo fatorial com dois fatores e estimação *maximum likelihood*, CP–1c=modelo com um componente principal, CP–2c=modelo com dois componentes principais, *r*=correlação estimada entre fatores (ou componentes) da solução, com base na rotação *oblimin*

Evidentemente, existem métodos exploratórios mais apropriados para determinar a dimensionalidade de um conjunto de dados (Golino & Epskamp, 2017; Lorenzo-Seva et al., 2011; Ruscio & Roche, 2012; Timmerman & Lorenzo-Seva, 2011). Não se recomenda, em hipótese alguma, que uma análise exploratória seja conduzida cegamente – sem hipóteses teóricas e sem o auxílio de uma ou mais dessas técnicas de investigação de dimensionalidade. A principal mensagem do presente editorial é outra. O alerta é que, se a dimensionalidade é desconhecida e o pesquisador comete *overfactoring*, suas chances de fazer uma escolha inapropriada de itens é maior se a técnica de base for a análise de componentes principais. Por isso, se o modelo teórico é reflexivo, isto é, se ele especifica uma conexão causal das variáveis latentes para os itens, a análise fatorial deve ser a técnica de escolha. No entanto, a análise de componentes principais pode ser útil quando o objetivo for puramente a redução do número de variáveis, sem a preocupação com um construto latente.

Referências

Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., & Hair Jr., J. F. (2012). *Análise multivariada de dados* (6th ed.). Bookman.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd Edition). The Guilford Press.

Damásio, B. F. (2012). Uso da análise fatorial exploratória em psicologia. *Avaliação Psicológica, 11*(2), 213-228.

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLOS ONE, 12*(6), e0174035. doi: 10.1371/journal.pone.0174035

Gorsuch, R. L. (1990). Common Factor Analysis versus Component Analysis: Some Well and Little Known Facts. *Multivariate Behavioral Research, 25*(1), 33-39. doi: 10.1207/s15327906mbr2501_3

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The Hull Method for Selecting the Number of Common Factors. *Multivariate Behavioral Research, 46*(2), 340-364. doi: 10.1080/00273171.2011.564527

Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning (Multivariate Applications Series)*. Routledge.

Revelle, W., & Revelle, M. (2015). Package 'psych.' In *The Comprehensive R Archive Network.*
Rosseel, Y. (2012). lavan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2).
Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24*(2), 282-292. doi: 10.1037/a0025697
Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (Fifth). Pearson.
Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220. doi: 10.1037/a0023353
Yanai, H., & Ichikawa, M. (2006). Factor Analysis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 257-296). Elsevier B.V. doi: 10.1016/S0169-7161(06)26009-7

Como citar este artigo

Hauck-Filho, N., & Valentini, F. (2020). A análise de componentes principais é útil para selecionar bons itens quando a dimensionalidade dos dados é desconhecida? [Editorial]. *Avaliação Psicológica, 19*(4), A-C. http://dx.doi.org/10.15689/ap.2020.1904.ed