Editor Científico: Francisco Antonio Serralvo
ISSN 2178-0080
PUC-SP
RAD Management in Dialog Review
Revista Administração em Diálogo
24 ANOS RAD
Avaliação: *Double Blind Review* https://doi.org/10.23925/2178-0080.2022v24i3.54115

Construção e validação de instrumentos de pesquisa de *Survey*: da psicologia à administração

Construction and Validation of Survey Research Instruments: From Psychology to Administration

Kelmara Mendes Vieira[I]
Aureliano Angel Bressan[2]

Resumo

Trata-se de um estudo descritivo e prescritivo, que apresenta os procedimentos teóricos, experimentais e analíticos necessários para a construção ou adaptação de um instrumento de pesquisa. O artigo traça um caminho para a construção de instrumentos, indicando para os pesquisadores quais os principais procedimentos, técnicas e cuidados que precisam ser adotados tanto na fase de construção teórica quanto na etapa de validação empírica. Especifica ainda o objetivo e os critérios a serem utilizados na aplicação da Análise Fatorial Exploratória, da Teoria da Resposta ao Item e da Modelagem de Equações Estruturais, no contexto da validação. Ao final, são apresentadas sugestões aos pesquisadores em administração que precisam construir ou adaptar instrumentos para pesquisas de *Survey*.

Palavras-chave: instrumentos de pesquisa, validade, *surveys*, psicometria

Abstract

This descriptive and prescriptive study presents the theoretical, experimental, and analytical procedures necessary for the construction or adaptation of a research instrument. The article traces a path for constructing instruments, indicating to researchers which main procedures, techniques, and precautions to adopt, both in the theoretical and empirical validation phases. It also specifies the objective and criteria for applying Exploratory Factor Analysis, Item Response Theory, and Structural Equation Modeling in the validation context. To build or adapt instruments for survey research, suggestions to management researchers were presented at the end.

Keywords: *research instruments, validity, surveys, psychometrics*

[1] kelmara@terra.com.br, Universidade Federal de Santa Maria - PPG Adm Pública, Santa Maria/RS [Brasil].
[2] aureliano.bressan@gmail.com, Universidade Federal de Minas Gerais - CEPEAD/UFMG, Belo Horizonte/MG [Brasil].

Recebido em: 06/05/2021
Aprovado em: 14/04/2022

Introdução

No mundo acadêmico, grande parte das pesquisas quantitativas são *surveys*. E o sucesso de uma *survey* invariavelmente está condicionado à construção de um bom instrumento de pesquisa. Entretanto, muitos pesquisadores não dão a devida atenção ou não se esforçam para dar ao instrumento o rigor científico necessário para que o mesmo seja capaz de atender ao objetivo da pesquisa.

Percebe-se, nas pesquisas publicadas em diferentes áreas do conhecimento, grandes diferenças quanto a este rigor metodológico. Na área de Administração, por exemplo, ainda há pouca discussão sobre a prática de adaptação das escalas (Heggestad et al. 2019) sendo comum o artigo informar que usou uma escala "validada" por algum autor, sem uma discussão mais ampla de quais os procedimentos de validação teriam sido aplicados. Neste contexto, uma simplificação comum é o uso de escalas construídas em outro país, utilizando-se apenas da técnica de tradução reversa, embora a adaptação de instrumentos, do ponto de vista psicométrico, vá muito além da simples tradução. Por outro lado, em áreas como a psicologia, observa-se um cuidado metodológico maior quando o assunto é a construção de medidas.

A medição é fundamental na ciência, e sem dúvida as qualidades mais importantes de uma medição são a validade e a confiabilidade (Clark; Watson, 2019). Segundo a 5ª edição do Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing [AERA, APA, & NCME], 2014) a validade se refere ao grau em que as evidências e a teoria suportam as interpretações do teste para os usos propostos do mesmo, representando de modo preciso a característica que se pretende medir. Além da medição e da validade, a fidedignidade ou confiabilidade é outro componente fundamental da qualidade de um instrumento. Esta propriedade está relacionada à capacidade do mesmo de reproduzir resultados de forma consistente ao longo do tempo (Souza; Alexandre; Guirardello, 2017). É um critério fundamental da qualidade de um instrumento (de Andrade Martins, 2006).

Assim, é neste contexto que este artigo se insere. O objetivo é apresentar aos pesquisadores em Administração um caminho para a construção de um instrumento de pesquisa, baseado na tradição da Psicometria.

Entende-se que este artigo apresenta uma grande utilidade prática, especialmente para os pesquisadores ainda incipientes na temática de construção e validação de instrumentos de pesquisa em Administração. Ter um caminho delineado, indicando o que precisa ser feito e como pode ser realizado, costuma ser um bom começo para aqueles que precisam se dedicar ao tema.

O artigo é estruturado então da seguinte maneira. A seção seguinte apresenta a proposta metodológica que delineará o caminho a ser seguido. A terceira seção especifica os procedimentos teóricos necessários nas fases iniciais da construção de um instrumento. A quarta seção é dedicada aos procedimentos experimentais. A 5ª seção apresenta os procedimentos analíticos. A sexta seção traz as considerações finais sobre os usos desta proposta metodológica.

PUC-SP
Revista Administração em Diálogo
ISSN 2178-0080
Programa de Pós-Graduação em Administração
Pontifícia Universidade Católica de São Paulo

O modelo de construção e validação de instrumentos

Este trabalho caracteriza-se como um estudo descritivo e prescritivo que se propõe a apresentar ao pesquisador da área de Administração um caminho para a construção e validação de instrumentos de pesquisa. Utiliza o aporte teórico da psicometria e inspira-se no modelo de elaboração de instrumentos proposto por Pasquali (2010). Este modelo estabelece três grupos de procedimentos: teóricos, empíricos e analíticos.

Os procedimentos teóricos, descritos na seção 3, enfocam a questão da fundamentação teórica do construto ou construtos para os quais se pretende desenvolver um instrumento de medida, bem como a operacionalização do construto em itens. Discute ainda os procedimentos para a análise semântica e de juízes, que compõem a validação de conteúdo dos itens do instrumento.

Os procedimentos empíricos ou experimentais definem as etapas e técnicas de aplicação do instrumento piloto e da coleta dos dados. São apresentados diversos aspectos sobre os quais o pesquisador deve se preocupar ao planejar o instrumento, sua aplicação e a coleta dos dados.

Por fim na seção cinco, os procedimentos analíticos, envolvendo as três técnicas estatísticas que podem ser utilizadas para a construção de um instrumento válido, e se for o caso, normatizado. Nesta seção também são feitas recomendações sobre o uso das técnicas e a indicação de softwares, com indicação de literaturas mais avançadas para o aprofundamento em cada uma das técnicas.

Procedimentos teóricos

Definição do Construto e Dimensões

A descrição do construto deve envolver seu contexto teórico e seu contexto hierárquico-estrutural, incluindo seu nível de abstração e como ele se distingue de construtos próximos (*near-neighbor constructs*) (Clark; Watson, 2019).Assim, a partir de uma base teórica sólida, geralmente obtida de amp las e sistemáticas revisões de literatura, além de consultas a peritos da área e na própria experiência pessoal, o pesquisador deve definir claramente o que o construto é e o que ele não é. Também é necessário estabelecer como a medida proposta para o construto se diferencia teórica ou empírica das medidas existentes (Shaffer; Deggest; Li, 2016), ou que lacuna da medição o estudo propõe a preencher.

Para Pasquali (2010) a definição do construto deve incluir dois tipos de definições. A definição constitutiva, a qual aparece tipicamente tal como a definição de termos em dicionários, e que caracteriza o construto, dando as dimensões que ele deve assumir no espaço semântico da teoria na qual está inserido. Em paralelo, deve ser feita ainda a definição operacional, onde o construto é definido em termos de operações concretas, ou seja, comportamentos por meio dos quais o construto se expressa.

Desenvolvimento do Conjunto de Itens

Após a definição do construto, o pesquisador deve construir os itens. Segundo Pasquali (2010) a construção dos itens deve seguir doze critérios, a saber: comportamental; objetividade, simplicidade, clareza, relevância, precisão, variedade,

PUC-SP
Revista Administração em Diálogo
ISSN 2178-0080
Programa de Pós-Graduação em Administração
Pontifícia Universidade Católica de São Paulo

modalidade, tipicidade, credibilidade, amplitude e equilíbrio. Por outro lado, Clark e Watson (2019) afirmam que deve-se evitar, neste processo, (a) o uso de expressões que podem ficar datadas rapidamente, (b) o uso de linguagem coloquial, que pode não ser familiar entre gerações, grupos étnicos, regiões ou gêneros, (c) itens que qualquer pessoa (p.ex. "às vezes estou mais feliz do que em outros momentos") ou que ninguém vai confirmar (p.ex. "Estou sempre furioso"), e (d) itens complexos ou de duplo sentido, que acessem mais de uma característica; por exemplo "Eu nunca beberia e dirigiria por poder ser parado pela polícia".

Outra questão importante e que merece especial atenção no desenvolvimento dos itens é o tipo de escala (Zickar, 2020). As escalas mais comuns são as dicotômicas (sim/não, verdadeiro/falso) e as escalas tipo Likert, usadas com diferentes significados, como por exemplo: concordância (concordo fortemente a discordo fortemente), frequência (p.ex., nunca a sempre), similaridade (muito similar a mim a nada similar a mim), e grau ou alcance (p.ex., nem um pouco a muito) (Comrey, 1988).

Além da escolha do significado da escala, outra questão importante é o número de pontos. A escala original de Likert (1932) inclui 5 pontos, mas o número de pontos varia enormemente entre as pesquisas. Por exemplo Simms, Zelazny, Williams, e Bernstein, (2019), avaliando escalas de concordância de 2 a 11 pontos, sugerem o uso de seis pontos. Em geral, o número de pontos mais utilizados varia de cinco a sete (Pasquali, 2010), sendo escalas com menos de 4 pontos menos recomendadas. A escolha do número de pontos deve levar em consideração também o número de itens do instrumento, já que muitos itens com muitos pontos implicarão necessariamente num aumento significativo da quantidade de alternativas disponíveis, e consequentemente, exigem do entrevistado maior esforço na tomada de decisão.

Além do cuidado com a construção de cada item e de suas escalas, cabe ao pesquisador a verificação do conjunto de itens que formam o instrumento. Loevinger (1957) argumenta que "the items of the pool should be chosen so as to sample all possible contents which might comprise the putative trait according to all known alternative theories of the trait" (p. 659). Já Segundo Clark e Watson (2019), duas implicações importantes deste princípio são as de que: (i) o instrumento inicial deve ser mais amplo e abrangente do que a visão teórica do construto alvo e, (ii) deve incluir conteúdo que acabará sendo eliminado no instrumento final. Na prática, o que se quer dizer aqui é que a definição do conjunto de itens deve levar em consideração o fato de que as análises psicométricas posteriores serão capazes de indicar quais itens podem de fato ser retirados da medição, mas serão incapazes de indicar quais deveriam ser incluídos. Logo, nessa etapa é preciso garantir que o conjunto de itens formulados sejam capazes de abarcar todos os aspectos da definição dos construtos.

Após a construção do conjunto de itens, é fundamental que o pesquisador realize a etapa de análise de conteúdo, pois esta representa o início de mecanismos para associar conceitos abstratos com indicadores observáveis e mensuráveis (Alexandre; Coluci, 2011) A análise de conteúdo, por vezes dividida em análise semântica e de juízes, visa avaliar o grau em que cada elemento de um instrumento de medida é relevante e representativo de um construto (Haynes; Richard; Kubany, 1995). Ou seja, a mesma avalia em que proporção os itens selecionados para medir uma construção teórica representam adequadamente todas as dimensões do conceito a ser medido.

O objetivo da análise semântica é o de avaliar a compreensão dos itens para toda população alvo (Pasquali, 2010). Neste aspecto, recomenda-se um estudo com indivíduos com a mesma característica da população, mas com diferentes níveis e áreas de formação, perfis de rendas e idade. Entende-se por indivíduos com a mesma característica da população aqueles que se enquadram no objeto de estudo do instrumento. Por exemplo, se o instrumento pretende medir stress em enfermeiros, a análise semântica deve necessariamente incluir enfermeiros com diferentes perfis.

Nesta etapa, é preciso que o pesquisador esteja consciente de que, por ser um especialista no tema, seu nível de conhecimento e compreensão de determinadas expressões, processos, procedimentos e outros vocábulos inerentes à temática pode ser muito diferente do apresentado pelo público-alvo. Daí a necessidade de que a análise semântica seja realizada com públicos de diferentes perfis, visando garantir que o instrumento seja compreensível a todos os entrevistados.

Após as adaptações das questões sugeridas na fase de análise semântica, o instrumento deve ser submetido à análise de juízes com diferentes especialidades. Estes especialistas devem avaliar o instrumento, ou mais especificamente, cada escala quanto à dimensão representada pelo item; ao grau de relevância do item e; quanto à adequação da formulação do item.

A literatura apresenta controvérsias sobre qual o número ideal de especialistas. Lynn (1986) recomenda um mínimo de cinco e um máximo de dez pessoas participando desse processo. Outros autores sugerem de seis a vinte sujeitos, sendo composto por um mínimo de três indivíduos em cada grupo de profissionais selecionados para participar (Haynes; Richard; Kubany, 1995). Nessa decisão, deve-se levar em conta as características do instrumento, a formação, a qualificação e a disponibilidade dos profissionais (Lynn, 1986, 7). Indica-se também a inclusão de pessoas leigas potencialmente relacionadas com a população do estudo (Rubio et al., 2003). A inclusão de pessoas leigas asseguraria a correção de expressões e termos que não estão muito claros. Quando for um processo de adaptação cultural, sugere-se a formação de um comitê multidisciplinar. Para que a análise de juízes seja conduzida de forma adequada são recomendados vários procedimentos, a saber:

1) Realizar um convite formal aos membros do comitê de juízes com uma carta explicando porque o sujeito foi escolhido como juiz e a relevância dos conceitos envolvidos e do instrumento como um todo. Recomenda-se também incluir o objetivo do estudo, as definições conceituais que deram origem ao instrumento, as dimensões envolvidas e o modelo de medida usado. Pode ainda incluir informações sobre o contexto e a população envolvida.

2) Construir as orientações específicas para cada etapa de avaliação. A primeira etapa deve ser de avaliação da especificação dos domínios, onde os juízes devem avaliar o instrumento como um todo, ou seja, se cada domínio ou conceito foi adequadamente coberto pelo conjunto de itens e se todas as dimensões foram incluídas. Na segunda, analisar os itens individualmente verificando sua clareza e pertinência.

3) Construir um instrumento específico para a avaliação dos especialistas. Geralmente emprega-se uma escala tipo Likert para fins de avaliação do instrumento. A Tabela 01 sugere algumas alternativas de perguntas e escalas.

Revista Administração em Diálogo
ISSN 2178-0080
Programa de Pós-Graduação em Administração
Pontifícia Universidade Católica de São Paulo

Tabela 1
Sugestões de dimensões e escalas a serem utilizadas no instrumento para avaliação dos juízes.

Critério de Análise	**Descrição para o juiz**	**Questão de avaliação do instrumento**	**Categorias de Resposta**
Dimensão/ Construto*	Analise a adequação de cada item em representar as dimensões/construtos da teoria estudada.	O Item pertence a qual dimensão/construto?	Inserir uma categoria para cada dimensão/construto, além da opção "outra: especifique"
Adequação da formulação*	Considerando as características da amostra a ser investigada, avalie a adequação do item no que se refere à linguagem, clareza e objetividade na proposição do conteúdo.	A formulação do item está adequada?	Sim, Não: Sugestão de aperfeiçoamento do item.
Grau de pertinência*	Considere se cada item foi elaborado de forma a avaliar o conceito de interesse.	O item é pertinente?	1-Nada Pertinente, 2-Pouco Pertinente, 3-Pertinente, 4-Muito Pertinente
Grau de relevância*	Considere o grau de associação entre o item e o construto. Analise o quanto o item é relevante para medir o construto	O item é relevante?	1-Nada Relevante, 2-Pouco Relevante, 3-Relevante, 4-Muito Relevante
Escala	Todos os itens do instrumento apresentarão uma escala tipo *Likert* de "especificar a escala" (exemplo: condordância:1-Discordo Totalmente,2- Discordo, 3-Indiferente,4- Concordo, 5-Condordo Totalmente	A escala é adequada?	Sim, Não: Sugestão de aperfeiçoamento/ mudança da escala.
Instruções do Instrumento	Descreva quais serão as instruções apresentadas antes da apresentação dos itens. Exemplo: A seguir apresentaremos uma lista de afirmações sobre "nome da escala". Marque com um "X" segundo seu grau de concordância com cada uma delas. Fique tranquilo ao responder as frases, pois não existem respostas certas ou erradas.	As instruções são adequadas?	Sim, Não: Sugestão de aperfeiçoamento/ mudança nas instruções

* questões que devem ser respondidas para cada item do instrumento
Fonte: elaborada pelos autores

Sugere-se que, no documento a ser enviado aos juízes, seja inserida uma seção denominada "Introdução" na qual o pesquisador apresentará as definições especificadas na coluna "descrição" da Tabela 01. Neste espaço inicial deve-se também apresentar as definições de todas as dimensões/construtos para que o pesquisador tenha o conhecimento necessário para identificar a qual dimensão o item pertence. Em seguida sugere-se que, para cada item do instrumento, sejam elaboradas as quatro questões relativas á dimensão, formulação, pertinência e relevância. E, no final sejam

PUC-SP
Revista Administração em Diálogo
ISSN 2178-0080
Programa de Pós-Graduação em Administração
Pontifícia Universidade Católica de São Paulo

inseridas as questões relativas à escala e instruções. Caso o instrumento utilize diferentes escalas, deve-se apresentar os questionamentos para cada escala, geralmente ao final do conjunto de itens que as utilizará.

No caso de adaptação cultural, a avaliação deve assegurar que a versão final seja totalmente compreensível e avaliar a sua equivalência cultural. Há a necessidade de existir equivalências semântica, idiomática, conceitual e experimental. Equivalência semântica é relativa ao significado das palavras (vocabulário, gramática); equivalência idiomática refere-se às expressões idiomáticas e coloquiais; equivalência experimental aborda situações coerentes com o contexto cultural; e, finalmente, a equivalência conceitual se refere ao conceito explorado (Guillemin, 1995). Nestes casos, devem ser apresentadas no instrumento questões para que os juízes avaliem cada uma destas equivalências.

Após a obtenção das respostas, inicia-se a fase quantitativa da análise conteúdo. Nesta etapa, para avaliar o nível de concordância entre os juízes pode se calcular o Coeficiente de Validade de conteúdo (CVC) e o Kappa. O CVC mede a proporção ou porcentagem de juízes que estão em concordância sobre determinados aspectos do instrumento e de seus itens (Alexandre; Coluci, 2011). O ponto de corte recomendado nesse tipo de análise é o mínimo de 0,80 sendo desejável 0,90 (Polit, Beck, 2006)

Para a análise de concordância da dimensão teórica de cada item entre os juízes e entre cada juiz e as dimensões pré-estabelecidas pelos autores da escala indica-se o cálculo do Kappa. O coeficiente Kappa de Cohen é a razão da proporção de vezes que os juízes concordam (corrigido por concordância devido ao acaso) com a proporção máxima de vezes que os juízes poderiam concordar (corrigida por concordância devido ao acaso). É aplicável quando os dados são categóricos e estão em uma escala nominal (Conger, 2017) (Warrens, 2020). Sendo que valores abaixo de 0,40 representam baixa concordância, valores situados entre 0,40 e 0,75 representam concordância mediana e valores maiores que 0,75 representam excelente concordância.

Ao final desta etapa finalmente o pesquisador terá em mãos todos os dados necessários para a construção do instrumento que será levado à campo. A partir da análise de conteúdo estarão definidos então os itens, as escalas e as instruções. Em seguida, o pesquisador poderá iniciar os procedimentos experimentais.

Procedimentos experimentais

Os procedimentos experimentais envolvem o planejamento, a aplicação e a coleta dos dados (Pasquali, 2010), os quais representam a fase inicial da validação do instrumento. Na fase de planejamento define-se a estrutura do instrumento e a amostra. No que se refere ao instrumento, Cizek (2020) especifica uma série de detalhamentos que pesquisador deve definir (Tabela 02).

Tabela 02
Aspectos do instrumento que podem ser definidos pelo pesquisador

Categoria	Descrição	Exemplos
Modo de apresentação	Especificar o modo pelo qual as direções, questões, tarefas ou solicitações do teste são apresentadas.	Papel ou interface web Texto, áudio ou apresentação em vídeo Tipo de fonte, tamanho e outras características de apresentação Instruções do teste ou leitura das questões
Modo de resposta	Especificar o modo pelo qual o participante fornece as respostas às questões, tarefas ou solicitações do teste.	Resposta em tempo real versus resposta gravada Uso de um pesquisador assistente para registrar as respostas
Programação do teste	Estabelecer o mês, dia ou intervalo de tempo em que o teste deve ser realizado.	Data fixa versus agendamento de teste sob demanda Janelas temporais durante as quais o teste deve ser realizado entre datas específicas Período do dia (p.ex., manhã ou tarde) Ordem fixa ou flexível de aplicação das seções do teste
Configuração do teste	A configuração do espaço físico no qual o participante irá realizar o teste precisa ser especificada.	Configurações de acomodação Orientação da tela Espaçamento Administração coletiva ou individual Variações permitidas na configuração (em termos de estímulos sonoros ou visuais) Luminosidade, temperatura e ambiente sugeridos Materiais proibidos na sala de testes (p.ex., gráficos, cartazes)
Tempo de teste	O tempo disponível para os participantes completarem o teste.	Tempo-limite especificado vs. possibilidade de extensão de tempo. Possibilidade de intervalos, frequência e duração dos intervalos entre seções do teste ou sob demanda.
Ferramentas de assistência	Especificar, caso necessário, auxílios que os participantes podem solicitar durante o teste.	Uso de dicionários, glossários ou manuais Uso de ferramentas de apoio (calculadoras, teclado ou tela *touch-screen*)

Fonte: adaptado de Cizek (2020).

Feito o planejamento e partindo do pressuposto que o instrumento foi construído para uma certa população, esta deve ser claramente definida e delimitada em termos de suas características específicas. Conforme Pasquali (2010), é necessário que o pesquisador determine qual é o tipo de indivíduo, em termos de características sociodemográficas, que constitui a população meta do estudo.

Então, a partir da especificação desta população meta, estima-se a amostra a ser estudada. Para a definição amostral podem ser utilizados critérios amostrais estatísticos

como tamanho mínimo amostral, para determinado nível de confiança e erro, e aleatoriedade na seleção dos participantes. Mas, em estudos psicométricos também são sugeridas regras de bolso, tais como: 1) se estiver seguro sobre quantos fatores o instrumento mede, um mínimo de 100 sujeitos por fator; 2) se houver dúvidas quanto ao número de dimensões, estimar uma amostra de 10 sujeitos para cada item do instrumento; e 3) dificilmente amostras com menos de 200 sujeitos serão consideradas adequadas para aplicação de análise fatorial e TRI (Pasquali, 2010). Os procedimentos experimentais se encerram com a coleta dos dados, devendo ser seguidos os quesitos estabelecidos na fase de planejamento. Em seguida, tem-se a fase de aplicação do instrumento, na qual o pesquisador deve decidir se a aplicação será presencial ou online. Ambas as formas possuem vantagens e desvantagens que podem impactar na qualidade dos dados (Triga; Manavopoulo, 2019) (de Boni, 2020). Ressalta-se que, no caso em que os instrumentos são aplicados presencialmente, é indispensável o treinamento adequado dos aplicadores visando garantir a padronização das condições de aplicação do instrumento. E, nos casos de pesquisas online, deve-se observar os procedimentos recomenddos pela literatura (para maiores detalhes veja Hewson; Vogel; Laurent, 2016; Zhang et al. 2017, Hewson, 2017, Faran; Zanbar, 2019)

Procedimentos analíticos

Após a coleta dos dados, inicia-se a etapa da validação do instrumento, a qual inclui os procedimentos analíticos. Para esta fase podem ser utilizadas diversas técnicas estatísticas, dentre as quais destacam a análise fatorial exploratória, a teoria da resposta ao item e a modelagem de equações estruturais.

Análise Fatorial Exploratória

A análise fatorial exploratória (AFE) é um dos procedimentos mais utilizados no desenvolvimento, avaliação e refinamento de instrumentos. A AFE tem como objetivo encontrar a estrutura subjacente em uma matriz de dados e determinar o número e a natureza das variáveis latentes (fatores) que melhor representam um conjunto de variáveis observadas (Brown, 2015). As variáveis observadas pertencem a um mesmo fator quando, e se, elas partilham uma variância em comum.

Um pré-requisito para a AFE é a existência de correlações entre as variáveis. As matrizes mais utilizadas são as correlações de Pearson, tetracórica e policórica. Correlações tetracóricas e policóricas, em comparação ao coeficiente de Pearson, tendem a ser uma estimativa mais consistente da verdadeira relação linear entre variáveis que possuem respostas nominais, ordinais ou do tipo Likert. Por isso, AFE com o uso de estimadores que se valem de correlações policóricas tendem a acertar com mais frequência o número de fatores subjacentes aos dados, produzindo ainda estimativas paramétricas mais consistentes de cargas fatoriais e correlações entre fatores (Asún; Rdz-Navarro; Alvarado, 2015).

O primeiro passo para a utilização da AFEs é observar se a matriz de dados é passível de fatoração. Para isso, dois métodos de avaliação são mais comumente utilizados: o critério de Kaiser Meyer-Olkin (KMO); e o Teste de Esfericidade de Bartlett. Os testes devem ser utilizados para indicar o grau de suscetibilidade ou ajuste dos dados à análise fatorial, ou seja, a mensuração do nível de confiança quando do tratamento dos dados por este método multivariado (HAIR et al., 2014). O teste de

esfericidade de Barlett é aplicado com a finalidade de avaliar se a correlação entre as variáveis é significativa, garantindo que apenas alguns fatores sejam capazes de representar grande parte da variabilidade dos dados. Valores do teste com níveis de significância $p < 0,05$ indicam que a matriz é fatorável (Tabachnick; Fidell, 2019).

Já o teste de KMO, representa a razão da correlação ao quadrado para a correlação parcial ao quadrado entre as variáveis e apresenta valores normalizados entre 0 e 1 (Field, 2017). Como regra para interpretação dos índices de KMO, valores menores que 0,5 são considerados inaceitáveis, valores entre 0,5 e 0,7 são considerados medíocres; valores entre 0,7 e 0,8 são considerados bons; valores maiores que 0,8 e 0,9 são considerados ótimos e excelentes, respectivamente (Hutcheson; Sofroniou, 1999).

Sendo a matriz de dados fatorável, deve-se então decidir pelo método de estimação. A literatura indica vários métodos para a estimação das cargas fatoriais, dentro os quais se destacam: Análise dos Componentes Principais (PCA), Análise Fatorial Principal (PFA), máxima verossimilhança, principais eixos fatoriais, mínimos quadrados generalizados, mínimos quadrados não ponderados e fatoração alfa. A literatura recomenda que, de início, seja utilizada a Análise Fatorial Principal como método de extração das cargas fatoriais (Comrey; 1988; Clark; Watson, 2019).

Existe a possibilidade de encontrar tantos fatores quantas forem as variáveis pesquisadas. Contudo, geralmente o pesquisador busca reduzir as informações contidas nas variáveis originais em um número menor de fatores. Assim, diversos procedimentos e critérios de retenção fatorial foram desenvolvidos, como por exemplo: a determinação a priori; o uso de autovalores, gráfico de declive, percentagem de variância, confiabilidade meio a meio, testes de significância e análise paralela. Sugere-se a utilização da análise paralela (Horn, 1965) com a permutação aleatória dos dados observados (Parallel Analysis with random permutation of observed data) (TIMMERMAN; LORENZO-SEVA, 2011) para verificar a dimensionalidade da escala.

A análise paralela (AP) é um procedimento estatístico de simulação de Monte-Carlo que consiste na construção aleatória de um conjunto hipotético de matrizes de correlação de variáveis, utilizando como base a mesma dimensionalidade do conjunto de dados reais. O número de fatores nos dados reais a ser retido refere-se àqueles que apresentam autovalor > 1 e que apresentam valor maior do que o respectivo autovalor obtido por meio dos dados aleatórios (O'Connor, 2000). Para aumentar a acurácia do método, deve-se considerar o intervalo de confiança de 95% obtido nos valores dos autovalores aleatórios (Patil et at., 2018).

Considerando que nem sempre os fatores escolhidos são de fácil interpretação, recorre-se à técnica de rotação dos eixos. Seguindo recomendação de Watson (2012), indica-se estimar os resultados tanto das rotações ortogonais como das oblíquas e avaliar qual resultado apresenta maior consistência com a construção teórica.

Para a seleção dos itens que permanecerão na análise, Clark e Watson (2019) recomendam a eliminação de itens com: (i) cargas primárias abaixo de 0,35 a 0,40 (para escalas mais amplas, e 0,45 a 0,50 para escalas de menor dimensão), e (ii) que tenham cargas similares ou maiores em outros fatores, embora estas diretrizes possam ser relaxadas em algumas circunstâncias.

A consistência interna dos fatores pode ser avaliada pelo Alfa de Cronbach (Cronbach, 1951) e pelo Ômega de McDonald (ω - Mcdonald, 1999). para os quais valores iguais ou superiores a 0,7 são considerados adequados (Hair et al., 2014).

Uma sugestão de software para a estimação da análise fatorial com dados categóricos é o programa Factor 10.10.01 (Lorenzo-Seva; Ferrando, 2017). Também é possível utilizar o pacote psych do ambiente R. Para exemplos de aplicação da AFE na validação de construtos veja Macêdo e Silva (2020), Dedeoğlu et al. (2020)

Teoria da Resposta ao Item

A Teoria de Resposta ao Item (TRI) tem ganhado espaço na literatura relacionada ao desenvolvimento de escalas. O método é baseado na pressuposição de que as respostas ao item refletem níveis de um construto subjacente e, além disso, que a relação resposta-característica de cada item pode ser descrita por uma função crescente monotônica chamada curva característica do item (CCI). Assim, indivíduos com maiores níveis da característica possuem maior probabilidade de responder ao item em uma determinada direção. Por exemplo, em uma escala de conhecimento, com respostas certas ou erradas, indivíduos com maiores habilidades possuem maior probabilidade de acertar o item Portanto, na TRI a ênfase recai na identificação de itens específicos que são os mais informativos para cada indivíduo, dado o seu nível na dimensão subjacente (Clark; Watson, 2019).

Para instrumentos considerados unidimensionais, são utilizados modelos que pressupõem a existência de uma única dimensão principal, os quais podem ser modelos logísticos (para os casos de escalas com uma única resposta correta) ou modelos politômicos (para os casos com escala do tipo Likert). Dentre os modelos unidimensionais geralmente os mais utilizados na construção de instrumentos, tem-se o modelo Rasch para escala logística e o Modelo de Resposta Gradual (MRG) de Samejima (1969), para escalas politômicas. Já nos casos em que o instrumento é multidimensional, ou seja, possui mais de um construto latente, recomenda-se o uso dos modelos multidimensionais de resposta ao item (Multidimensional item response theory -MIRT)

Os modelos da TRI permitem estimar parâmetros diferentes dos itens, em especial, a probabilidade de acerto ao acaso, a discriminação e a dificuldade. O modelo mais utilizado para análise de itens na TRI é o modelo logístico. Para ilustrar, vamos considerar itens dicotômicos. Neste caso, há basicamente três especificações, que se diferenciam pelo número de parâmetros que utilizam para descrever o item: (i) modelos logísticos de 1 parâmetro, que avaliam somente e dificuldade do item, (ii) modelos logísticos de 2 parâmetros, que avaliam a dificuldade e a discriminação do item, e (iii) modelos logísticos de 3 parâmetros, que avaliam a dificuldade, a discriminação e a probabilidade de resposta correta para um respondente com baixa habilidade.

Para o modelo de 3 parâmetros, a especificação é a seguinte:

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_j)}}$$

Com i = 1, 2, ..., I e j = 1, 2, ..., n, em que:

Uij é uma variável dicotômica que assume o valor 1 quando o respondente j responde corretamente ao item i, e 0 quando o indivíduo j não responde corretamente a este i-ésimo item;

θj representa a habilidade ou traço latente do j-ésimo respondente;

P(Uij = 1 | θj) é a Função de Resposta ao Item (FRI). A mesma corresponde à probabilidade de um respondente j com habilidade θj responder corretamente ao item i.

bi é o parâmetro de dificuldade (ou de posição) do item i, medido na mesma escala de habilidade.

ai é o parâmetro de discriminação (ou de inclinação) do item i, com valor proporcional à inclinação da Curva de Característica do Item (CCI) no ponto bi.

ci é o parâmetro do item que representa a probabilidade de indivíduos com baixa habilidade responderem corretamente o item i. Também conhecida como probabilidade de acerto ao acaso.

D é um fator de escala, constante e igual a 1.

Todavia, não existe um valor exato de ai para decidir se um item discrimina bem ou não. Baker (2001) estabelece uma tabela de classificação em que valores de 0,01 a 0,34 são muito baixos, de 0,35 a 0,64 são baixos, 0,65 a 1,34 são moderados, 1,35 a 1,69 são altos e maiores que 1,70 são muito altos.

Já o parâmetro bi de dificuldade do item indica que, quanto maior seu valor, mais difícil é o item e, portanto, indivíduos com habilidade alta terão uma maior probabilidade de acertá-lo em relação a indivíduos com baixa habilidade. A relação entre a resposta prevista do item e o traço latente do indivíduo é conhecida através da Curva Característica do Item (CCI).

Pode-se ainda, analisar a quantidade de informação que um item fornece para a medida do traço latente. A Função de Informação do Item (FII) avalia a informação fornecida por cada item e reflete a qualidade do mesmo, permitindo analisar quanto um item contém de informação para a medida de habilidade. Também pode-se estimar a Função de Informação do Teste (FIT) a qual é simplesmente a soma das informações fornecidas por cada item que compõe o instrumento.

O uso da TRI oferece duas vantagens importantes sobre outras estratégias de seleção de itens (Clark; Watson, 2019). Primeiro, o método permite a especificação do nível de características em que cada item do instrumento é mais informativo. Essas informações podem ser usadas para identificar um conjunto de itens que produzem avaliações precisas, confiáveis e válidas em todo o intervalo da característica. Assim, os métodos de TRI oferecem uma capacidade aprimorada de discriminar entre indivíduos nos extremos das distribuições de características (por exemplo, entre aqueles com habilidade muito alta e muito baixa em determinado traço). Segundo, os métodos de TRI permitem a estimativa do nível de característica de cada indivíduo sem a necessidade de administrar um conjunto fixo de itens. Essa flexibilidade permite o

desenvolvimento de testes adaptativos utilizando algoritmos de combinação de itens (Clark; Watson, 2019)

Para estimação, pode-se utilizar o programa R. Neste, o pacote Irtoys (Partchev, 2016) pode ser aplicado para a estimação dos modelos logísticos unidimensionais como o Rasch e o pacote ltm (Rizopoulos, 2015) para os modelos politômicos. Já para modelos multidimensionais, o pacote mirt versão 1.30 (Chalmers, 2012) possui uma generalização do modelo de Samejima (196) para itens politômicos, podendo-se utilizar o algoritmo de extração Metropolis-Hastings Robbins-Monro (Mhrm – CAI, 2010) e a integração quasi-Monte Carlo. Alguns exemplos de aplicação da teoria da resposta ao item na validação de construtos são encontrados em Tezza et al. (2018), Vieira; Potrich e Bressan (2020) Vieira et al. (2020), e Houts e Knoll (2020).

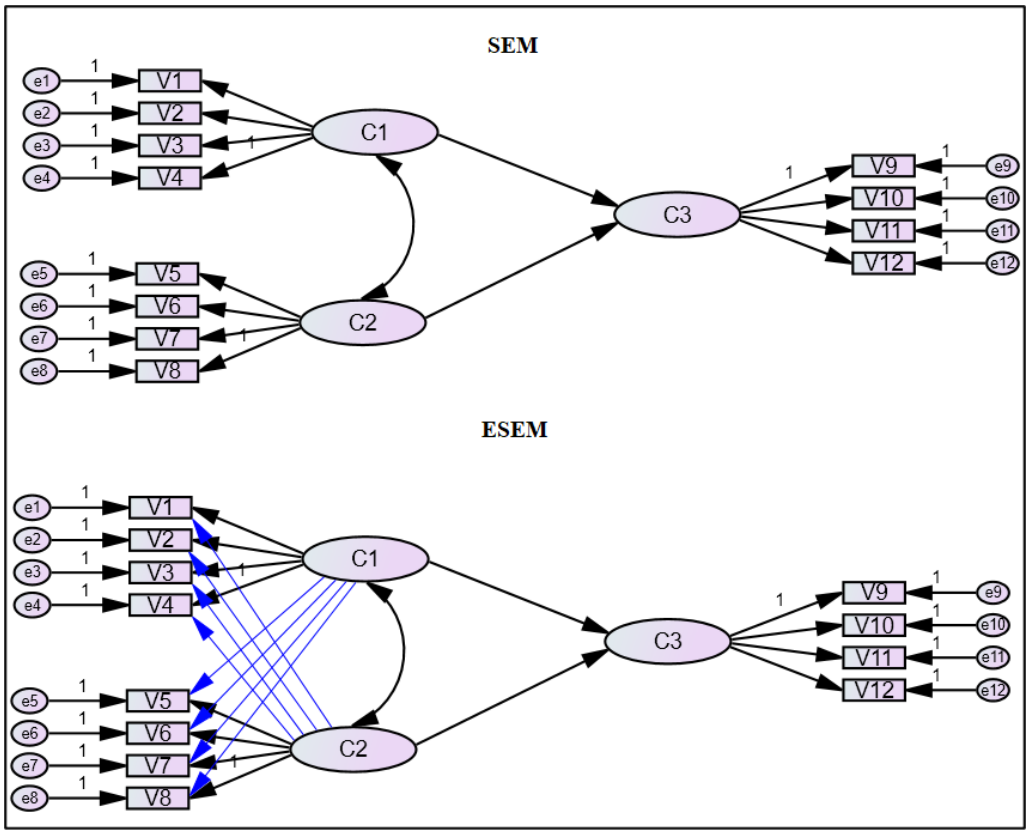
Modelagem de Equações Estruturais.

A modelagem de equações estruturais (SEM-Structural Equation Modeling) é uma técnica apropriada para avaliar a validade de construto. A validade de construto se refere ao grau em que um instrumento está medindo o construto de interesse (Souza, Alexandre & Guirardello, 2017).

Numa etapa inicial de validação de um instrumento, geralmente se utiliza a análise fatorial exploratória, conforme descrita anteriormente, com os objetivos de: identificar o número de construtos (unidimensional ou multidimensional) e avaliar a importância de manter ou retirar itens. Já a SEM avança na busca da validação ao examinar as relações teóricas dos itens do instrumento e os conceitos definidos na teoria e por promover evidências a respeito das relações hipotéticas entre os construtos (Waltz; Strickland; Lenz, 2016).

Na SEM o pesquisador assume que a variável latente (construto) tem uma estrutura fatorial concisa, sem cargas fatoriais cruzadas, e avalia o modelo de mensuração através da análise fatorial confirmatória (CFA- *confirmatory fator analysis*). Entretanto, quando se deseja avaliar a existência de cargas fatoriais cruzadas, ou seja, quando um item do instrumento pode estar teoricamente relacionado a mais de um construto, o pesquisador pode optar pela modelagem de equações estruturais exploratória (ESEM- *exploratory structural equation modeling*) cujo modelo de mensuração da variável latente usa a análise fatorial exploratória ao invés da CFA (Mai; Zhang; Wen, 2018). A Figura 1 exemplifica a diferença entre a SEM e a ESEM.

Figura 1
Exemplos da estrutura de modelos SEM e ESEM



Fonte: elaborada pelos autores.

Na Figura 1 os retângulos representam as variáveis ou itens (V), as elipses menores os erros de mensuração das variáveis (e) e as maiores representam os construtos (C). Na SEM os pesquisadores indicam, a partir da construção teórica, a qual construto cada variável do instrumento pertence, o que torna o modelo de mensuração mais simples. Enquanto na ESEM cada item pode ser alocado a mais de um construto, como nos casos das variáveis V1 a V8, permitindo uma maior investigação da estrutura de cada construto, ou seja, quais itens são mais bem alocados em quais construtos.

Nestas modelagens a avaliação dos modelos é tradicionalmente realizada a partir dos índices de ajuste e da significância das cargas fatoriais. Os índices mais comumente utilizados são $\chi 2/df$, *comparative fit index* (CFI), índice de Tucker-Lewis (TLI), raiz do erro quadrado médio da aproximação (RMSEA), erro quadrado médio padronizado residual (SRMR), critério de informação de Akaike (AIC), e Critério de informação Bayesiano (BICF).

Para o $\chi 2/df$, RMSEA, SRMR, AIC, e BIC, quanto menor a estatística, melhor o ajuste, ao passo que para o CFI e para o TLI, valores maiores indicam melhor qualidade de ajuste. De acordo com Hu e Bentler (1999), as regras de decisão para ajuste dos modelos consideram que o ajuste é adequado quando $\chi 2=df < 5$, CFI > .95, TLI > .95 RMSEA < .06, e SRMR < .08. Os critérios AIC e BIC são usados para comparar modelos aninhados ou não-aninhados.

Para avaliação do modelo de mensuração, é comum verificar as validades convergente e discriminante. A validade convergente pode ser avaliada através da comparação da pontuação do teste que está m consideração e a pontuação de um outro teste que teoricamente mede o mesmo construto. Neste caso, a existência de uma correlação forte e positiva entre os resultados dos dois testes fornecem evidências de validade convergente. Na validade convergente os itens indicadores de um construto possuem uma elevada proporção de variância em comum. Já a validade discriminante é o grau em que um construto se difere dos demais (Hair et al., 2014, Byrne, 2016).

Para avaliar a validade convergente pode-se utilizar a magnitude das cargas fatoriais, a Variâncias Média Extraída (AVE) e a Confiabilidade Composta. A literatura indica que as cargas fatoriais devem ser de pelo menos 0,5 ou superiores. Se um item apresentar valores inferiores a 0,5 torna-se um forte candidato a deixar o modelo fatorial. A AVE verifica a proporção da variância dos itens que são explicados pelo construto ao qual pertencem. Valores de AVE iguais ou superiores a 0,5 sugerem que o modelo converge (Hair et al., 2014). Já para a validade discriminante, utiliza-se a análise das cargas cruzadas e a comparação das raízes quadradas da AVE. Os itens devem apresentar cargas fatoriais mais elevadas nos construtos que foram previamente designados do que nos demais, e as raízes quadradas das AVE devem ser maiores do que a correlação entre os construtos (Hair et al., 2014).

Existem na literatura diversos softwares para a estimação da modelagem de equações estruturais: Amos, SAS PROC CALIS, e os pacotes no ambiente R sem, lavaan, OpenMx, LISREL, EQS, e Mplus (veja Narayanan, 2012, para uma discussão sobre os programas). Dentre os métodos de estimação, o Método de Mínimos Quadrados Ponderados e Ajustados por Variância (Weighted Least-Squares Mean and Variance-Adjusted) com matrizes de correção policóricas é o mais utilizado. Para alguns exemplos de aplicação da modelagem na validação de modelos Potrich, Vieira, Mendes-da-Silva, (2016), Abrantes-Braga e Veludo-De-Oliveira (2019), , Cruz, Ferreira e Gabardo-Martins (2020), Dos Reis Soares et al. (2020).

Normatização

Padronização ou normatização, em seu sentido mais geral, refere-se à necessidade de existir uniformidade em todos os procedimentos no uso de um teste válido e preciso. Engloba desde as precauções a serem tomadas na aplicação do teste (uniformidade das condições de testagem, controle do grupo, instruções padronizadas, etc) até o desenvolvimento de parâmetros e critérios para a interpretação dos resultados (Pasquali, 2017). Mais especificamente, Cronbach e Neto (1996) distinguiu a padronização como sendo a uniformidade na aplicação dos testes e a normatização como a uniformidade na interpretação dos escores dos testes.

No caso das pesquisas na área de administração, na maioria dos casos, o pesquisador atua na adaptação de instrumentos e, como tal, deverá também adaptar as instruções e normas estabelecidas na escala original. Já nos casos em que o pesquisador se propõe a construir uma nova escala, então todo processo de padronização e normatização deverá ser criado seguindo os critérios psicométricos.

Considerações finais

Avançar nas técnicas de construção de um instrumento é uma necessidade eminente para todos os pesquisadores que trabalham ou desejam trabalhar com pesquisas envolvendo surveys. Mesmo que o objetivo da pesquisa não seja a construção de uma escala, o uso de alguns critérios de adaptação e validação são necessários para que possa garantir a qualidade do instrumento e consequentemente a possibilidade de que se consiga medir o que realmente se pretende. Afinal, uma boa pesquisa quantitativa requer o uso de medidas bem construídas e validadas (Devellis, 2016), mesmo que um bom instrumento sozinho não garanta o sucesso do estudo.

Ao longo dos anos, tem-se visto o uso frequente em pesquisas a argumentação de que a escala foi "validada" ou "adaptada, termos amplos frequentemente utilizados para indicar que os autores mudaram algo na escala (Heggestad et al., 2019), mas muitas vezes sem a verificação e explicitação de quais os procedimentos foram realizados. E, em muitos casos, observa-se que os estudos citados na argumentação muitas vezes fizeram uso de uma tradução reversa e do cálculo da consistência interna pelo Alpha de Cronbach, pois também não tinham como objeto principal a validação ou construção do instrumento. Entende-se, todavia, que o pesquisador deve ter todo o cuidado necessário ao usar essas expressões na argumentação do uso de suas escalas, deixando-as exclusivamente para os casos em observe que a maioria dos procedimentos de validação tenham sido realizados de modo adequado.

Uma vez que o pesquisador decida utilizar uma escala construída em outras línguas e ainda não adaptada ao Brasil, sugere-se a realização dos procedimentos apresentados neste artigo. Entretanto, sabe-se que, nem sempre a pesquisa tem como objetivo principal a construção ou validação de um instrumento, e que o tempo e os recursos disponíveis podem não ser suficientes para a realização de todas as etapas. Neste caso, sugere-se que, caso exista um outro instrumento ou escala já validado para o mesmo tema, o pesquisador pode optar pela utilização de outra escala já validada. Caso não existam escalas validadas, entende-se que o pesquisador deve buscar cumprir pelo menos todos os procedimentos teóricos aqui discutidos.

Já se a opção for pela construção de uma escala, é preciso que o pesquisador tenha em mente que o processo é cíclico, ou seja, nem sempre ao final dos procedimentos sugeridos tem-se uma escala pronta. Geralmente, este processo interativo envolve um ciclo inicial de desenvolvimento do instrumento, coleta de dados e avaliação psicométrica, seguida de um ciclo adicional de revisão do instrumento de medida e do construto subjacente. Cabe ressaltar ainda que diferentes versões de uma medida podem ser construídas, incluindo aí versões reduzidas, traduções ou adaptações para grupos específicos que também requerem o uso de diversos processos psicométricos.

Ressalta-se, por fim, que os caminhos aqui delineados não são únicos, existindo na literatura, especialmente de psicometria, outros procedimentos e técnicas que podem e devem ser praticados por pesquisadores interessados na temática. Nosso objetivo principal foi apresentar um dos caminhos possíveis, o qual apresenta os pontos principais que podem ser seguidos pelos pesquisadores em Administração.

Referências

Abrantes-Braga, F. D. M., & Veludo-de-Oliveira, T. (2019). Development and validation of financial well-being related scales. *International Journal of Bank Marketing*.

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

Asún, R. A., Rdz-Navarro, K., & Alvarado, J. M. (2016). Developing multidimensional Likert scales using item factor analysis: The case of four-point items. *Sociological Methods & Research*, 45(1), 109-133.

Baker, Frank B. (2001). The basics of item response theory. Recuperado de http://files.eric.ed.gov/fulltext/ED458219.pdf.

Boni, R. B. D. (2020). Websurveys en tiempos de la COVID-19. *Cadernos de Saúde Pública*, 36, e00155820.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3 ed). Routledge

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3), 307-335.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, 48(6), 1-29.

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, 56(5), 754.

Conger, A. J. (2017). Kappa and rater accuracy: Paradigms and parameters. *Educational and Psychological Measurement*, 77(6), 1019-1047.

Costa Alexandre, N. M., & Orpinelli Coluci, M. Z. (2011). Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. *Revista Ciência & Saúde Coletiva*, 16(7).

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

Cronbach, L. J., Neto, C. A. S., & Veronese, M. A. V. (1996). *Fundamentos da testagem psicológica*. Artes Médicas.

Cruz, R. P. D. S., Ferreira, M. C., & Gabardo-Martins, L. M. D. (2020). Evidências de validade para a escala de alegria no trabalho. *Revista Psicologia Organizações e Trabalho*, 20(1), 941-946.

de Andrade Martins, G. (2006). Sobre confiabillidade e validade. *Revista Brasileira de Gestão de Negócios-RBGN*, 8(20), 1-12.

Dedeoğlu, B. B., Taheri, B., Okumus, F., & Gannon, M. (2020). Understanding the importance that consumers attach to social media sharing (ISMS): Scale development and validation. *Tourism Management*, 76, 103954.

DeVellis, R. F. (2016). Scale development: Theory and applications (Vol. 26). Sage publications.

Faran, Y., & Zanbar, L. (2019). Do required fields in online surveys in the social sciences impair reliability?. *International Journal of Social Research Methodology*, 22(6), 637-649.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, 29(2), 236-240.

Field, A. (2017). *Discovering statistics using IBM SPSS statistics*. North American edition. Sage.

Guillemin, F. (1995). Cross-cultural adaptation and validation of heatth status measures. *Scandinavian Journal of Rheumatology*, 24(2), 61-63.

Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications.

Hair, Joseph F. et al. (2014) *Multivariate data analysis*: *Pearson new international edition*. Pearson Education Limited.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, 7(3), 238.

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Monroe Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management*, 45(6), 2596-2627.

Hewson C., Vogel C., Laurent D. (2016). *Internet research methods*. Sage Publishing.

Hewson, C. (2017). Research design and tools for online research. *The Sage Handbook of Online Research Methods*, 57-75.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179-185.

Houts, C. R., & Knoll, M. A. (2020). The financial knowledge scale: New analyses, findings, and development of a short form. *Journal of Consumer Affairs*, 54(2), 775-800.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal*, 6(1), 1-55.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage.

Likert, R. (1932). *A technique for the measurement of attitudes*. Archives of psychology.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3), 635-694.

Lynn, M. R. (1986). *Determination and quantification of content validity*. Nursing research.

Macêdo, J. W. D. L., & Silva, A. B. D. (2020). Construção e Validação de uma Escala de Competências Socioemocionais no Brasil. *Revista Psicologia Organizações e Trabalho*, 20(2), 965-973.

Mai, Y., Zhang, Z., & Wen, Z. (2018). Comparing exploratory structural equation modeling and existing approaches for multiple regression with latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(5), 737-749.

Mcdonald, R. P. (1999). *Test theory: a unified treatment*. Lawrence Earlbaum Associates. Inc., Mahwah, NJ, pp. 142-145.

Narayanan, A. (2012). A review of eight software packages for structural equation modeling. *The American Statistician*, 66(2), 129-138.

O'connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32(3), 396-402.

Partchev, I (2016). *Package irtoys: Simple interface to the estimation and plotting of IRT models*. CRAN R Project. http://cran.r-project.org/web/packages/irtoys/irtoys.pdf.

Pasquali, L. (2009). *Instrumentação psicológica: fundamentos e práticas*. Artmed Editora.

Pasquali, L. (2017). *Psicometria: teoria dos testes na psicologia e na educação*. Editora Vozes Limitada.

Patil, Vivek H. et al. (2007) *Parallel analysis engine to aid determining number of factors to retain* (Computer software). https://analytics.gonzaga.edu/parallelengine/

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489-497.

Potrich, A. C. G., Vieira, K. M., & Mendes-Da-Silva, W. (2016). Development of a financial literacy model for university students. *Management Research Review*, 39(3), 356-376.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1-25.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, 27(2), 94-104.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2), 100.

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods*, 19(1), 80-110.

Siegel, S., & Castellan Jr, N. J. (1975). *Estatística não-paramétrica para ciências do comportamento*. Artmed Editora.

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, 31(4), 557.

Soares, F. H. D. R., Carvalho, A. V. D., Keegan, E., Neufeld, C. B., & Mansur-Alves, M. (2020). Adaptação e validação da escala de perfeccionismo Almost Perfect Scale-Revised para o português brasileiro. *Revista Avaliação Psicológica*, 19(3), 310-321.

Souza, A. C. D., Alexandre, N. M. C., & Guirardello, E. D. B. (2017). Psychometric properties in instruments evaluation of reliability and validity. *Epidemiologia e Serviços de Saúde*, 26, 649-659.

Tabachnick, B. G., Fidell, L. S. (2019). *Using multivariate statistics* (Vol. 7, pp. 481-498). Boston, MA: Pearson.

Tezza, R., Bornia, A. C., Andrade, D. F. D., & Barbetta, P. A. (2018). Modelo multidimensional para mensurar qualidade em website de e-commerce utilizando a teoria da resposta ao item. *Gestão & Produção*, 25(4), 916-934.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, 16(2), 209.

Triga, V., & Manavopoulos, V. (2019, March). Does mode of administration impact on quality of data? Comparing a traditional survey versus an online survey via a Voting Advice Application. In *Survey Research Methods* (Vol. 13, No. 2, pp. 181-194).

Vieira, K. M., Martins, P. S. R., Bender Filho, R., & Júnior, F. D. J. M. (2020). Escala de Determinantes da Evasão no Ensino a Distância (EDED): Proposição e Validação. *EaD em Foco*, 10(2).

Vieira, K. M., Potrich, A. C. G., & Bressan, A. A. (2020). A proposal of a financial knowledge scale based on item response theory. *Journal of Behavioral and Experimental Finance*, 28, 100405.

Waltz, Carolyn F.; Strickland, Ora Lea; Lenz, Elizabeth R. (Ed.) (2016). *Measurement in nursing and health research*. Springer Publishing Company.

Warrens, M. J. (2020). Kappa coefficients for dichotomous-nominal classifications. *Advances in Data Analysis and Classification*, 1-16.

Watson, D. (2012). Objective tests as instruments of psychological theory and research. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 349–369). American Psychological Association. https://doi.org/10.1037/13619-019

Zhang, X., Kuchinke, L., Woud, M. L., Velten, J., & Margraf, J. (2017). Survey method matters: Online/offline questionnaires and face-to-face or telephone interviews differ. *Computers in Human Behavior*, 71, 172-180.

Zickar, M. J. (2020). Measurement development and evaluation. Annual Review of Organizational. *Psychology and Organizational Behavior*, 7, 213-232.

PUC-SP
Revista Administração em Diálogo
ISSN 2178-0080
Programa de Pós-Graduação em Administração
Pontifícia Universidade Católica de São Paulo