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

Construindo escalas de autorrelato: O que fazer?

Ariela Raissa Lima-Costa[1]
Universidade São Francisco, Campinas-SP, Brasil
Bruno Bonfá-Araujo
University of Western Ontario, Londres-ON, Canadá

RESUMO

Escalas de autorrelato são comuns no cotidiano de pesquisadores, porém detalhes, que às vezes parecem de pouca importância, acabam sendo desconsiderados. Autorrelato são caracterizados pela pessoa de interesse ser a própria fonte de informação do pesquisador, assim a forma em que as possibilidades de resposta são apresentadas é de suma importância. Os formatos de resposta mais comuns são tipo Likert e as escolhas forçadas que vem ganhando popularidade nos últimos anos. Neste artigo são apresentados os elementos necessários que devem ser observados quando se constrói ou se escolhe escalas que possuam esses dois formatos de resposta. São discutidos também a quantidade de itens ou blocos necessários, as opções de resposta, polaridade dos itens, sistema de correção, bem como ao final são indicadas boas práticas na construção de itens para ambos os formatos. Dessa forma, pretende-se contribuir para a Psicologia, em especial a área de construção de instrumentos de autorrelato.

Palavras-chaves: Autorrelato; Questionários; Testes Psicológicos.

ABSTRACT – Developing self-report scales: what to do?

Self-report scales are common in the daily lives of researchers, however, details, which sometimes seem of little importance, can end up being disregarded. Self-reports are characterized by the person of interest being the researcher's own source of information, therefore the way in which the response possibilities are presented is of paramount importance. The most common response formats are Likert type and forced choice, which have been gaining popularity in recent years. This article presents the elements that must be observed when developing or choosing scales that have these two response formats. Also discussed are the number of items or blocks needed, the answer options, item polarity, the correction system, and good practices in the construction of items for both formats. The intention is to contribute to Psychology, especially the area of construction of self-report instruments.

Keywords: Self Report; Questionnaires; Psychological Tests.

RESUMEN – Construcción de escalas de autoinforme: ¿Qué hacer?

Las escalas de autoinforme son habituales en el día a día de los investigadores, pero los detalles, que a veces parecen de poca importancia, acaban siendo desestimados. Los autoinformes se caracterizan porque la persona de interés es la propia fuente de información del investigador, por lo que la forma en que se presentan las posibilidades de respuesta es de suma importancia. Los formatos de respuesta más comunes son tipo Likert y las elecciones forzadas, que han ido ganando popularidad en los últimos años. Este artículo presenta los elementos necesarios que se deben observar al momento de construir o elegir escalas que tengan estos dos formatos de respuesta. También se discute la cantidad de ítems o bloques necesarios, las opciones de respuesta, la polaridad de los ítems, el sistema de corrección, así como las buenas prácticas en la construcción de ítems para ambos formatos. De esta forma, se pretende contribuir a la Psicología, especialmente al área de construcción de instrumentos de autoinforme.

Palabras clave: Autoinforme; Cuestionario; Tests Psicológicos.

As formas de coletar informações sobre a personalidade são diversas. Cattell (1958) distinguiu três modalidades, as quais ele denominou de registro de vida (*Life record observation* ou *L-data*), questionário (*Questionnaire data* ou *Q-data*) e testes objetivos (*Objective data* ou *T-data*). No primeiro, formato L-, o sujeito é avaliado por um ou mais observador externo, de modo que a investigação é feita no ambiente natural do sujeito, por meio de suas relações familiares e sociais e pode ser registrado, por exemplo, pela frequência que determinado comportamento aparece. O segundo, formato Q-, recolhe informações por meio da auto-observação, autoavaliação e autorrelato que o sujeito faz de si; são os questionários, inventários e escalas. O último, formato T-, avalia o sujeito por meio de sua reação a diferentes estímulos em um cenário controlado, usando estímulos aparentemente ambíguos, o que dificulta para a pessoa saber o que de fato está sendo avaliado, como testes situacionais e associações implícitas.

Cada formato de coleta determina o tipo de informação que deve ser acessada pelo avaliador. Tais

[1] Endereço para correspondência: Pós-graduação em Psicologia da Universidade São Francisco. Rua Waldemar César da Silveira, 105, Jardim Cura D'ars, 13045-510, Campinas, SP. *E-mail:* arielalima10@gmail.com

Artigo derivado da "O controle de desejabilidade social via diferentes formatos de resposta: avaliação da tríade sombria" de Ariela Raissa Lima-Costa com orientação de Nelson Hauck Filho, defendida em 2020 no programa de Pós-graduação em Psicologia com ênfase em Avaliação Psicológica da Universidade São Francisco.

informações estão distribuídas entre as que são acessíveis ao outro e/ou ao sujeito. Luft e Ingham (1961) elaboraram um esquema, denominado Janela de Johari, em que mostram quais informações estão disponíveis para avaliação direta ou indireta de pesquisadores e clínicos. Na Figura 1 são apresentados os quadrantes desenvolvidos por esses autores. No quadrante 'arena' as informações sobre a personalidade são percebidas pelo próprio sujeito e pelos outros, por exemplo uma pessoa extrovertida, que percebe em si e é visível para terceiros a facilidade para interagir com outras pessoas em diversos contextos.

No quadrante 'ponto cego', os outros observam algo que a própria pessoa não consegue perceber. Como exemplo, pode-se pensar em uma pessoa que considera seus comportamentos adequados a situação, mas para os outros ela é percebida como alguém hostil. No quadrante 'fachada', a pessoa, deliberadamente esconde informações sobre si, por exemplo, ela evita expressar pensamentos íntimos que ache que possa lhe causar embaraço. E por fim, o quadrante 'oculto', engloba o que é desconhecido, em que existem influências na forma de se comportar, pensar e sentir da pessoa que não são percebidas por ela nem pelos outros, por exemplo, uma pessoa que tem preconceito contra algum grupo e não demonstra isso para outros e não percebe em si (Luft & Ingham, 1961).

Figura 1
Modelo da Janela de Johari

	Conhecido pelo eu	Desconhecido pelo eu
Conhecido pelos outros	Arena	Ponto cego
Desconhecido pelos outros	Fachada	Oculto

Ao relacionar a proposta de Cattell (1958) e Luft e Ingham (1961) é possível identificar um formato de resposta para cada quadrante. As informações da 'arena' e 'ponto cego' podem ser acessadas por meio do formato L-, uma vez que este se baseia na observação externa. O formato Q- possibilita coletar informações disponíveis no quadrante 'arena' e 'fachada', posto que depende da percepção que o sujeito tem de si e de informações que, em sua maior parte, só podem ser acessadas por ele mesmo. O formato T- pode ser útil para acessar informações contidas no quadrante 'oculto', uma vez que o sujeito é avaliado por meio de estímulos aparentemente ambíguos que facilita que conteúdos desconhecidos ao sujeito se manifestem, e dificulta que responda baseado apenas em informações acessíveis diretamente, como testes de associação implícita.

O formato Q- (ou autorrelato, como será tratado no decorrer do texto) é muito utilizado na avaliação da personalidade (Chan, 2009) e o foco deste artigo. Isso ocorre, pois, o formato L- demanda mais tempo do pesquisador ou clínico e mais recursos pessoais e financeiros, pois o pesquisador vai a campo e passa um tempo amplo relatando os comportamentos da pessoa em diferentes situações, além de precisar entrar em contato com pessoas que fazem parte do círculo social daquela que está sendo avaliada. Do mesmo modo o formato T-, pois além da questão financeira e de tempo, ele é um método complexo tanto para a elaboração dos estímulos, como para a codificação e interpretação dos dados coletados (Cattell, 1958).

Além do tipo de informação que se pretende obter é preciso decidir de que forma o sujeito irá fornecê-las. No autorrelato, a forma de responder tem implicações da computação de escores e nas análises estatísticas usadas para tratar cada tipo de dado. O formato de resposta mais comum é o tipo Likert, porém outro formato vem ganhando atenção dos pesquisadores é o de escolha forçada (Weijters et al., 2010). Neste último, a pessoa é apresentada a dois ou mais itens e deve escolher o item que mais tem a ver consigo (Thurstone, 1928). No tipo Likert, são apresentadas categorias ordenadas de respostas, para que a pessoa indique a intensidade com que concorda ou discorda do conteúdo do item em relação a presença ou ausência daquela característica em si (Likert, 1932). As escalas tipo Likert surgiram com a ideia de simplificar o processo de resposta, já que a construção de escalas de escolha forçada demanda mais tempo para o pesquisador, pois há necessidade de criar mais itens e que tenham níveis similares de desejabilidade social e diferentes em termos de dificuldade.

São várias as decisões a serem tomadas na construção de escalas de autorrelato, seja no formato de escolha forçada ou escala Likert. Essas decisões influenciam no processo cognitivo envolvido no ato de responder a uma escala de autorrelato. Para responder a tais testes, o sujeito primeiro precisa compreender o que o item afirma (etapa de compreensão), o que implica em clareza de instrução, de escrita dos itens e o *design* do teste em si (Tourangeau et al., 2000), daí a importância dos elementos que serão apresentados a seguir. Para respeitar uma sequência histórica, primeiro serão apresentados os elementos de uma escala de escolha forçada, em seguida serão apresentados os elementos de uma escala com formato tipo Likert.

Elementos de escalas de escolha forçada

O formato de resposta de escolha forçada foi baseado na Lei de Julgamento Comparativo, proposto por Thurstone (1928). A ideia partiu da compreensão de que as atitudes estariam distribuídas em um contínuo com distâncias regulares, assim ao apresentar dois estímulos a pessoa tenderia a escolher o que julgasse estar mais próximo da forma que se percebe (i.e., utilidade do item). Dessa forma, para avaliar uma pessoa, seria preciso que os itens cobrissem todo o contínuo dos traços de interesse, isto é, a escala precisaria de itens que avaliassem características do extremo positivo ao negativo, com uma zona neutra ou intermediária (Thurstone, 1928). O método de escolha forçada aqui discutido é aquele apresentado por Brown (2014). As recomendações de construção de escalas em formato de escolha forçada sugerem que elas devem ser elaboradas considerando seis elementos: número de itens por bloco, a quantidade de blocos, tipos de opções de respostas, dimensionalidade dos blocos, polaridade dos itens e o sistema de correção (Brown & Maydeu-Olivares, 2011, 2016; Heggestad et al., 2006).

Número de itens

Os blocos devem conter no mínimo dois itens, os chamados duplets. O recomendado é que os itens tenham extensão curta, evitando negativas diretas, e apresentem o máximo de quatro itens por bloco, pois quanto maior a quantidade de itens maior será a complexidade cognitiva, além de cansar os sujeitos e contribuir para respostas aleatórias (Brown & Maydeu-Olivares, 2011).

Quantidade de blocos

A ideia de que mais itens, mais informação sobre os sujeitos é válida no formato de escolha forçada. No caso da escolha forçada, cada bloco é considerado um item e, assim, quanto maior a quantidade de blocos melhores informações serão obtidas. Brown e Maydeu-Olivares (2011) compararam 24 (12 blocos com 2 itens em cada) e 48 (24 blocos com 2 itens em cada) itens que avaliavam dois fatores, ambos forneceram medidas precisas dos valores esperados quando usados itens positivos e negativos, no modelo de 12 blocos a confiabilidade dos dados variou entre 0,645 e 0,756 e no modelo com 24 blocos a variação foi entre 0,812 e 0,877. Também compararam 60 itens que avaliavam cinco fatores divididos em 15 (4 itens por bloco), 20 (3 itens por bloco) e 30 blocos (2 itens por bloco), e verificaram que todos esses modelos fornecem medidas precisas, com itens positivos e negativos. Em outros estudos, foram comparados instrumentos com 18 (r=0,70-0,77) e 36 blocos, o último teve um melhor desempenho (r=0,73-0,83) em recuperar o escore verdadeiro (Hontangas et al., 2015; Hontangas et al., 2016). Assim, a escolha da quantidade de blocos depende da quantidade de traços avaliados e se os itens são positivos ou mistos (positivos e negativos).

Opções de respostas

Ao construir uma escala de formato de escolha forçada as opções de resposta são os formatos chamados PICK, MOLE e RANK (ver Figura 2). No formato PICK, termo que vem da palavra em inglês *"pick"* que significa "escolher", o sujeito deve escolher somente um item por bloco, sendo mais indicado para blocos com dois itens. No formato MOLE, termo que surgiu da junção da primeira sílaba de duas palavras em inglês *"most"* e *"least"*, que significam, respectivamente, "o mais" e "o menos", o sujeito indica em cada bloco qual item tem mais a ver consigo e o que tem menos a ver consigo, sendo indicado para blocos com três ou quatro itens. No formato RANK, termo que vem da palavra em inglês *"rank"* que significa "classificar", o sujeito deve classificar em ordem crescente o quanto cada item lhe descreve, sendo recomendado para blocos que contenham três ou mais itens. Em estudo de simulação, considerando o tamanho e discriminação do bloco, polaridade dos itens e variabilidade de parâmetros os modelos RANK (r=0,67-0,74), e MOLE (r=0,66-0,73) apresentaram desempenhos similares para estimar o escore verdadeiro, com uma diferença menor que 0,013 Hontangas et al., 2015; Hontangas et al., 2016).

Figura 2

Exemplo de tipos de opções de respostas em escalas de escolha forçada

PICK: Selecione o item que mais tem a ver com você.
a. Gosto de manipular as pessoas. b. Sou uma pessoa insensível.
MOLE: Selecione o item que MAIS tem a ver com você e o que MENOS tem a ver com você.
a. Gosto de manipular as pessoas. b. Sou uma pessoa insensível. c. Mereço tratamento especial dos outros.
RANK: Ordene os itens de 1 a 4, sendo que 1 indica o item que mais tem a ver com você e 4 o item que tem menos a ver com você.
a. Gosto de manipular as pessoas. ( ) b. Sou uma pessoa insensível. ( ) c. Mereço tratamento especial dos outros. ( ) d. Sinto alegria na maior parte do tempo. ( )

Dimensão do atributo avaliado

O atributo pode ser uni ou multidimensional. Os instrumentos multidimensionais têm blocos em que cada item se refere a uma dimensão independente, e que a distinção no conteúdo dos itens seja clara. Sendo sugerido que quanto maior o número de atributos e menor e/ou negativa for a correlação entre eles melhor a avaliação, pois a informação será influenciada majoritariamente por um traço (Brown & Maydeu-Olivares, 2011).

Polaridade dos itens

Dentro de cada bloco, a polaridade dos itens pode ser positiva, invertida (negativa) ou mista (positivos e negativos). Essa decisão está ligada à dimensionalidade do atributo avaliado. Os itens positivos avaliam a diferença entre dois traços, e itens invertidos a soma da influência de dois fatores, sendo que a combinação (itens mistos) deles permite localizar o item no contínuo do traço. Estudos de simulação sugerem que itens escritos em direções opostas (positivos e negativos) fornecem melhores dados (r=0,80-0,91) que itens apenas direcionados positivamente para estimar o escore verdadeiro (r=0,67-0,74) quando avaliada em relação a quantidade de itens, tipos de opções de resposta e dimensionalidade. Além disso, os itens precisam estar balanceados quanto ao nível de desejabilidade social, ou seja, é importante que tenham um nível valorativo semelhante, para evitar que a escolha de um item em detrimento do outro possa ser manipulada pela percepção que o respondente tem do que é mais socialmente esperado (Brown & Maydeu-Olivares, 2011; Hontangas et al., 2015; Hontangas et al., 2016).

Sistema de correção

Refere-se a como o escore de cada sujeito será contabilizado. Duas abordagens para corrigir escalas de escolha forçada foram identificados por Brown e Maydeu-Olivares (2016). O primeiro é baseado na abordagem clássica ou medida por "decreto", em que simplesmente somam-se os escores de cada item, isto é, escalas de escolha forçada são tratadas como escalas normativas o que faz com que sejam produzidos escores ipsativos ou quase-ipsativos, em que não há variabilidade na pontuação entre diferentes pessoas. O segundo é baseado em abordagens que usam a teoria de resposta ao item (TRI) como uma forma de contornar a formação de escores ipsativos, eles fazem uso de métodos de comparação binária de forma a possibilitar comparações interindividuais, pois permitem variação na pontuação das pessoas. Na Tabela 1 pode-se visualizar a correção de alguns itens com base na abordagem clássica e uma baseada em TRI.

A forma de correção clássica gera escores ipsativos. O termo 'ipsativo' foi usado inicialmente por Cattell para designar escores de uma escala que são constantes para todas as pessoas (Brown & Maydeu-Olivares, 2011; Brown & Maydeu-Olivares, 2013). Esse tipo de dado compromete a qualidade psicométrica da escala, uma vez que a pontuação em um atributo é dependente de outro atributo que está sendo avaliado, isto é, uma pessoa que pontua alto em um traço necessariamente vai pontuar baixo em outro, dessa forma a pessoa só pode ser comparada com ela mesma, a aplicação desse método pode ser visto na Tabela 1. Também distorce as evidências de validade do instrumento, por exemplo, devido à falta de variância a relação com outras escalas seja enviesada e dificulta a estimação da fidedignidade, pois viola os pressupostos de técnicas estatísticas usadas para calcular esse coeficiente, como a ideia de linearidade entre escore observado e escore verdadeiro no traço avaliado (Brown & Maydeu-Olivares, 2011; Brown & Maydeu-Olivares, 2013; Chan, 2003; Wang et al., 2017)

O sistema de correção Thurstonian baseado na teoria de resposta ao item apresenta a vantagem de poder ser usado tanto para instrumentos com estrutura unidimensionais e multidimensionais. Por meio dele é possível analisar blocos com qualquer quantidade de itens. Essa proposta de correção foi baseada na teoria de Thurstonian em 1931, em que ele dizia que uma pessoa ao responder a um teste comparava a utilidade (i.e., valor psicológico) de cada item com o seu nível individual no traço avaliado para decidir qual selecionar. Na prática são usadas comparações binárias entre pares de itens que formam um bloco. A quantidade de comparações obtidas é calculada pela fórmula $\tilde{n}=n(n-1)/2$ em que n corresponde ao número de itens por bloco, um exemplo de aplicação desse método pode ser visualizada na Tabela 1. O modelo estima carga fatorial, variância de erro, correlação entre os atributos, o valor de *thresholds* (probabilidade de 50% da pessoa selecionar uma categoria de resposta em detrimento a outra), e permite a formulação de um escore para cada sujeito (Brown & Maydeu-Olivares, 2012).

A Tabela 1 traz o exemplo de duas pessoas (A e B) que responderam ao mesmo teste e selecionaram diferentes itens, porém obtiveram o mesmo escore (coluna PC) em um item de escolha forçada com tipo MOLE de resposta. Na pontuação clássica, o item que mais tem a ver com o sujeito recebe 2 pontos, o que menos tem a ver recebe 0 e os outros recebem 1. Assim, independentemente dos itens selecionados e da pessoa a pontuação será sempre 3, a única comparação possível é entre a pessoa com ela mesma, isto é, um escore ipsativo. Quando o escore é calculado por uma abordagem baseada na TRI, por exemplo o modelo Thurstonian de TRI, em que se parte de combinações binárias, o item A é comparado com o item B e C, e o item B com o item C, se o primeiro item for selecionado em vez do segundo dá-se 1 ponto, em caso contrário dá-se 0. Elas permitem que haja variância nos escores e a comparação interpessoal se torna possível, logo fica perceptível que a Pessoa A tem mais traços da tríade sombria do que a Pessoa B. Assim, neste trabalho será considerado o sistema de correção Thustonian que se baseia na TRI.

Tabela 1
Comparação entre pontuação clássica (PC) e pontuação pelo modelo Thurstonian de TRI em um bloco de item de escolha forçada com tipo MOLE de opção de resposta

Itens	Pessoa A			
	Mais	Menos	PC	Thurstonian
a. Gosto de manipular as pessoas.	x		2	{A,B} = 0
b. Sou uma pessoa insensível.			1	{A,C} = 1
c. Considero-me melhor do que os outros.		x	0	{B,C} = 1
Total			3	2
Itens	Pessoa B			
	Mais	Menos	PC	Thurstonian
a. Gosto de manipular as pessoas.		x	0	{A,B} = 0
b. Sou uma pessoa insensível.	x		2	{A,C} = 0
c. Considero-me melhor do que os outros.			1	{B,C} = 1
Total			3	1

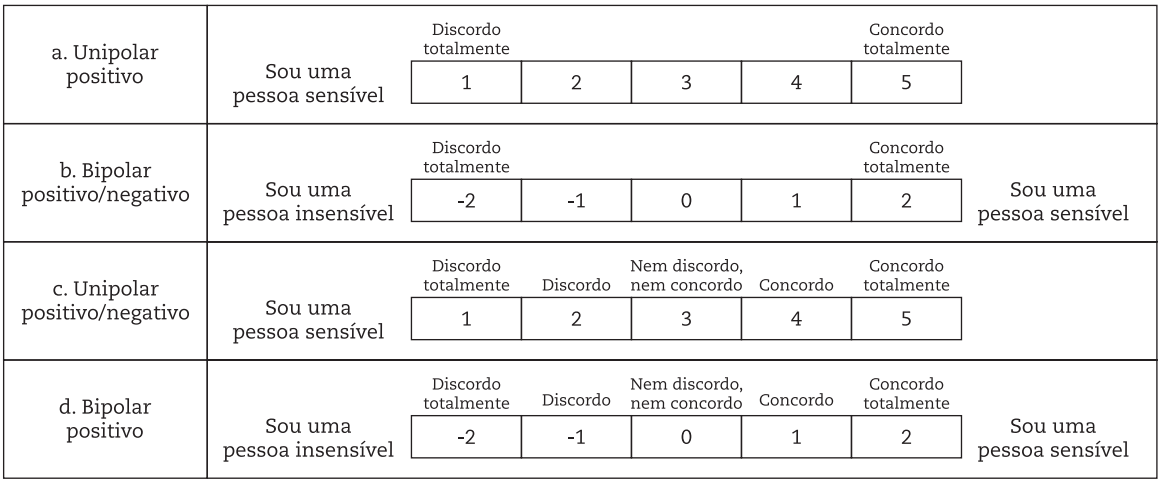
Elementos de escalas de formato tipo Likert

As escalas de resposta em formato tipo Likert são uma técnica não comparativa de escalonamento de respostas (Willits et al., 2016). Esse formato proporciona a avaliação absoluta de um item por vez, em que as pessoas indicam a valência (e.g., se concordam ou discordam) e a intensidade (e.g., pouco, moderado e muito) com que as afirmativas as descrevem. Esse formato foi inicialmente proposto por Likert (1932) que buscava uma forma de melhor compreender atitudes e comportamentos por meio de uma métrica comum e passível de tratamento estatístico, além de ser mais fácil de construir em comparação a proposta de Thurstone (1928). Na proposta inicial de Likert, a escala teria duas valências (desaprovação e aprovação) e cinco categorias de intensidade, em que 1 indicaria total desaprovação e 5 total aprovação. Apesar de aparentemente simples, esse formato de resposta requer que se considere a polaridade, numeração de categorias, rótulos, e quantidade de categorias.

Polaridade da escala

A escala pode ser unipolar (i.e., apresenta apenas um item em um extremo da escala) ou bipolar (i.e., apresenta um item em cada extremo da escala), como apresentado na Figura 3. Esse aspecto é explicado pelo efeito da simetria, em que os respondentes, ao se depararem com as categorias, percebem escalas bipolares como mais simétricas e mais fáceis de responder. No caso das escalas unipolares, elas têm o significado das categorias dividida de uma forma assimétrica, demandando um esforço cognitivo maior por parte do respondente para selecionar a categoria que melhor lhe representa (Cabooter et al., 2016).

Figura 3
Exemplos de formatos de escalas tipo Likert. Fonte: adaptado de Cabooter et al. (2016)



Numeração das categorias

Esta pode ser com todos os itens positivos e apresentados em ordem crescente (Figura 3, itens 'a' e 'd'), ou metade dos itens negativos e outra metade dos itens positivos (Figura 3, itens 'b' e 'c'). A influência desse aspecto pode ser explicada pelo efeito de intensidade, em que categorias extremas são percebidas como mais intensas do que as outras. A presença de itens positivos e negativos torna mais claro o significado dos pontos, principalmente dos pontos extremos por dar mais intensidade a eles, fazendo com que as pessoas escolham as categorias com mais cautela, o que pode contribuir para redução do viés de respostas extremas (Cabooter et al., 2016).

No estudo de Cabooter et al. (2016) eles verificaram a relação entre os tipos de polaridade quanto a efeito de simetria e intensidade, e a relação delas com viés de concordância e respostas extremas. Os resultados mostram que escalas unipolares positivas geraram menos concordância em relação aos outros três formatos, e menos respostas extremas em comparação com escalas unipolares positivas/negativas ($F(1330)=18{,}24$, $p<0{,}01$). As escalas bipolares positivas/negativas ($m=0{,}17$) geraram menos respostas extremas em relação às escalas bipolares positivas ($m=0{,}23$), possivelmente pelo efeito de intensidade que foi maior na primeira ($F(1330)=7{,}76$, $p<0{,}01$). Cada tipo apresentou limitações que podem interferir na distribuição de resposta, assim os autores recomendam que seja considerado o atributo que está sendo avaliado para decidir qual formato de escala se ajusta melhor.

Rótulos

Eles indicam o tipo de julgamento a ser feito pela pessoa (Parker et al., 2013), podendo indicar concordância, frequência, aprovação, performance, intensidade, qualidade, familiaridade, entre outros (Casper et al., 2019). Os mais comuns são os que vão desde "discordo totalmente" até "concordo totalmente", porém a escolha depende do objetivo da medida, do tipo de informação que se pretende obter (e.g., "nem um pouco familiarizado" até "extremamente familiarizado"). Mesmo que a simples tradução de categorias pareça algo comum, a intensidade percebida para um dado advérbio e/ou substantivo é diferente para cada cultura (e.g., o termo *'very'* do inglês pode ser traduzido como *'muito'* ou *'bastante'* e esses dois termos do português podem ser percebidos com intensidade variadas, em que alguns podem entender que muito > bastante ou bastante > muito). Consequentemente, isso pode influenciar na distribuição das respostas devido a diferenças na compreensão de intensidade dos rótulos (Dolnicar & Grün, 2013). O ideal é a realização de investigações com pessoas de cada cultura para entender a ordem de intensidade dada a advérbios como 'muito', 'bastante', 'extremamente', 'moderadamente', 'pouco', 'às vezes', 'sempre', entre outros que são encontrados em instrumentos de autorrelato que usam escalas de resposta tipo Likert (Weijters et al., 2013).

A posição dos rótulos é outro ponto importante. A dúvida recai sobre rotular todas as categorias de respostas (Figuras 3, itens 'c' e 'd') ou apenas os extremos (Figuras 3, itens 'a' e 'b'). Embora o assunto pareça de pouca importância, esse detalhe de construção da escala pode criar vieses de respostas, como investigado por Weijters, Cabooter e Schillewaert (2010). No estudo, os autores investigaram a influência de rótulos completos ou nos extremos em favorecer respostas aquiescentes (RA), extremas (RE) e respostas descuidadas a itens invertidos (RD). Por meio de análises de moderação e regressão, os resultados mostraram que escalas com rótulos completos tendiam a aumentar RA ($\beta=0{,}168$, $t=6{,}32$, $p<0{,}001$), diminuir RE ($\beta=0{,}436$, $t=-9{,}67$, $p<0{,}001$) e RD ($\beta=-0{,}490$, $t=-5{,}81$, $p<0{,}001$), pois a diferença entre categorias positivas e negativas é destacada, assim como destacado o ponto intermediário. Outro benefício é que assim se demanda um menor esforço cognitivo do respondente, já que o significado de todas as categorias está disponível. Por outro lado, o uso de rótulos apenas nos extremos teve valores consistentemente altos de variância explicada em todos os itens (entre 1,4 e 2,6). Isso sugere que esse formato proporciona melhores estimações em modelos lineares, isto é, em casos que se queira investigar e estimar a relação linear entre duas variáveis (e.g., correlações, regressões lineares; Weijters et al., 2010).

A familiarização com os rótulos é outro ponto a se considerar ao construir uma escala do tipo Likert. Quando a pessoa já respondeu testes de diversas temáticas, mas que usam o mesmo formato de resposta, no momento em que se depara com tal formato ela tende a interpretar as categorias (rótulos) baseadas em orientações passadas (Weijters et al., 2013). Isso pode fazer com que as pessoas respondam de forma descuidada, pois não há uma preocupação em avaliar o que está sendo afirmado antes de responder, como também pode acarretar informações equivocadas, pois o novo teste pode ter uma interpretação das categorias diferente das quais ela está habituada. Por exemplo, considerando escalas com cinco categorias, se a pessoa aprende que 1 indica discordância e 5 concordância ela tenderá a responder com essas interpretações, mas caso responda a um teste em que o pesquisador resolveu trocar para que 1 indique concordância e 5 discordância e a pessoa se abstém de ler as orientações, todas as repostas estarão erradas e a avaliação da pessoa apresentará outra interpretação.

Quantidade de categorias de resposta

Em um estudo comparativo Lee e Paek (2014) não identificaram diferenças nas propriedades psicométricas no uso de 4, 5 e 6 categorias, porém houve um decréscimo na qualidade quando consideradas 2 e 3 categorias. Weijters et al. (2010) verificaram que escalas de 5 pontos com rótulos completos tendem a diminuir respostas descuidadas para itens invertidos (RD), e que escalas com 5 ou 7 pontos com rótulos nos extremos estimam melhor

a relação linear entre variáveis. Weng (2004) investigou a relação entre a quantidade de categorias e o coeficiente alfa. Ele verificou uma relação positiva entre essas variáveis, sugerindo que aumentar a quantidade de categoria aumentaria o coeficiente alfa da escala. Não obstante, o autor destaca que a escolha dessa quantidade deve considerar a habilidade discriminativa da amostra investigada, pois quanto mais categorias mais esforço cognitivo é exigido dos respondentes.

A presença ou ausência do ponto neutro é outro aspecto a ser considerado na decisão da quantidade de categorias. A interpretação desse ponto varia e pode ser compreendido, segundo Baka e Figgou (2012), em três categorias: (a) falta de conhecimentos ou indiferença, quando a pessoa desconhece sobre o assunto investigado, quando tem dificuldade de compreender o conteúdo do item ou não se importa com o que está sendo avaliado; (b) dilema e ambivalência, essa interpretação é reflexo de uma verdadeira neutralidade do sujeito em relação ao que está sendo avaliado ou porque considera o sentido do que está sendo avaliado como eticamente conflitante; (c) oposição a afirmativa, quando a pessoa percebe o ponto neutro como uma forma de se opor ao que está sendo avaliado. Achados similares foram encontrados por Nadler et al. (2015), que perguntaram as pessoas qual o sentido do ponto neutro para elas, os significados mais frequentes foram não ter uma opinião (15%), não se importar (14%), incerteza da resposta (13%) e como sendo um ponto intermediário (10%). Assim, a decisão de usar um ponto neutro se relaciona a liberdade de escolha que o pesquisador disponibiliza ao respondente, se quer lhe dar uma opção de neutralidade ou forçá-lo a tomar uma decisão de concordância/discordância. Portanto, não há uma indicação fixa sobre seu uso, dependendo do objetivo e contexto da ferramenta a ser respondida (Chyung et al., 2017).

O mais comum é que tais decisões sejam tomadas de forma aleatória e sem uma preocupação com o impacto na qualidade da avaliação das pessoas. O desenho de escalas tipo Likert pode afetar na distribuição das respostas (i.e., vieses de resposta), na avaliação da homogeneidade ou heterogeneidade da amostra, na média, na correlação entre itens e na relação com variáveis demográficas, afetando a qualidade psicométrica e as conclusões geradas pelo escore do teste (Cabooter et al., 2016; Parker et al., 2013). Assim é importante considerar o construto que está sendo avaliado e a população a que se destina para verificar qual a melhor configuração de uma escala em que se pretende usar formato de resposta tipo Likert.

Boas práticas na construção de itens

Uma vez escolhido o formato (i.e., Likert *versus* escolha forçada) para ser utilizado no instrumento, bem como as especificidades do construto a ser mensurado, certas práticas podem facilitar o trabalho do pesquisador ao construir os itens. Na Tabela 2 são apresentados comparativos no que tangem alguns elementos de escalas de autorrelato.

Tabela 2
Boas práticas no processo de construção de itens

Recomendações	O que evitar?	Como pode ser feito?
Evite itens com conteúdo complexo	Sou lépido em festividades	Sou animado em festas
	Tenho tido uma conduta melancólica	Tenho me sentido triste
Evite itens com expressões regionais	Costumo ser o bode expiatório	Costumo ser culpado por coisas que não fiz
	Sou uma pessoa bolada	Sou uma pessoa preocupada
Evite itens com dupla negativas	Eu nunca falho com meus compromissos	Sou comprometido
	Eu não falho jamais com minhas atividades	Eu sempre faço minhas atividades
Evite itens com duas ideias e termos condicionais	Sou extrovertido e amável	Sou extrovertido
		Sou amável
	Sinto-me bem quando estou com meus colegas e meus familiares	Sinto-me bem com meus colegas
		Sinto-me bem com meus familiares
	Costumo ler ou sair com meus amigos para me divertir	Costumo ler para me divertir Costumo sair com meus amigos para me divertir
Evite itens com a palavra "não"	Não sou feliz	Sou infeliz
	Não sou organizado	Sou desorganizado

Ao propor um instrumento, os responsáveis pela pesquisa devem estar atentos e evitar certas práticas que tendem a exigir um maior esforço cognitivo dos respondentes, bem como influenciam negativamente as análises

realizadas após a coleta de dados (Artino et al., 2011; Furr, 2011). Itens com conteúdos complexos são inacessíveis a maior parte da população, bem como desmotivam os respondentes; itens bem construídos devem expressar de maneira clara e concisa o conteúdo. Expressões regionais podem representar uma escapatória para construtos de difícil definição, contudo pesquisadores devem estar cientes de que a maior parte destas não se aplicam a todos os grupos sociais e tornam o instrumento restrito à um pequeno grupo (*American Educational Research Association* [AERA], *American Psychological Association* [APA], & *National Council on Measurement in Education* [NCME], 2014).

Nessa mesma conformidade, itens com duplas negativas e que mensuram duas ideias são armadilhas que podem dificultar o processo de análises. Além de exigirem mais dos respondentes, o processo de mensuração se torna incerto, visto que as respostas atribuídas não podem ser consideradas totalmente fiéis, pois não fica evidente a que aspecto do item o sujeito deu mais ênfase ao responder, principalmente em casos de duas ideias e itens com termos condicionais. Por último, ao serem propostos instrumentos balanceados (i.e., a mesma quantidade de itens com direção positiva e negativa) para estimações estatísticas como a aquiescência; deve-se evitar apenas incluir a palavra "não" para inverter o conteúdo de um item. Dado que a inserção dessa palavra dificulta o processo de resposta com chaves que vão desde discordo totalmente até concordo totalmente (Furr, 2011; Hauck-Filho et al., 2021).

Os elementos apresentados são importantes no processo de construção de instrumentos de autorrelato, eles influenciam na quantidade e na qualidade de informação que os pesquisadores irão obter. Muitas vezes são tratados como aspectos secundários no processo de construção, porém como os estudos demonstraram cada elemento influencia na compreensão e julgamento que o sujeito faz antes de optar por uma resposta. Com isso, pretendeu-se deixar mais acessível esses conhecimentos e contribuir para a área de construção de instrumentos, especialmente, instrumentos que avaliam construtos psicológicos. Deve-se estar principalmente atento ao processo de construção dos itens, se estes retratam o construto de forma genuína, como também considerar práticas que visem a proporcionar uma melhor clareza e concisão para o respondente.

A preocupação com a qualidade do processo de construção de itens é tema recorrente na literatura brasileira. O Conselho Federal de Psicologia, por meio da Resolução 31/2022 (*Conselho Federal de Psicologia* [CFP], 2022) implementou padrões mais rigorosos na avaliação dos testes, incluindo análises que atestem a qualidade psicométrica dos itens que compõe instrumentos psicológicos, além da tradicional avaliação do escore total (Andrade & Valentini, 2018; Hauck-Filho, 2018). Assim, a avaliação da qualidade dos itens também está voltada para a minimização de vieses de resposta como aquiescência (Hauck-Filho et al., 2021; Valentini & Hauck-Filho, 2020) e desejabilidade social (Costa & Hauck Filho, 2017; Hauck-Filho & Valentini, 2019).

Este estudo não está isento de limitações, assim como não pretendeu esgotar a temática de desenvolvimento de instrumentos. Contudo, teve-se como principal recomendação colocar em foco uma prática, algumas vezes, menosprezada ao propor o desenvolvimento de uma ferramenta que auxilia os profissionais. É oportuno destacar tópicos não explorados aqui, como a apresentação visual dos itens, o processo de formatação de instrumentos, bem como a influência que tais escolhas podem ter (ver Artino & Gehlbach, 2012).

Agradecimentos

Agradecemos ao nosso orientador, Prof. Dr. Nelson Hauck Filho por seu constante acompanhamento e confiança em nosso trabalho, em especial na elaboração deste artigo, por nos possibilitar nos expressar como pesquisadores, mas sabendo que temos onde nos firmar para buscar apoio.

Financiamento

A pesquisa relatada no manuscrito foi financiada pela bolsa de doutorado da primeira autora. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Declaração de participação da elaboração do manuscrito

Declaramos que todos os autores participaram da elaboração do manuscrito. Especificamente, a autora Ariela Raissa Lima-Costa participou da redação inicial do estudo – conceitualização, investigação, visualização e o autor Bruno Bonfá-Araujo participou da redação final do trabalho – revisão e edição.

Disponibilidade dos dados e materiais

Todos os dados e sintaxes gerados e analisados durante esta pesquisa serão tratados com total sigilo devido às exigências do Comitê de Ética em Pesquisa em Seres Humanos. Porém, o conjunto de dados e sintaxes que apoiam as conclusões deste artigo estão disponíveis mediante razoável solicitação ao autor principal do estudo.

Conflito de interesses

Os autores declaram que não há conflitos de interesses.

Referências

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurment in Education [NCME] (2014). *Standards for educational and psychological testing.* Washington: American Educational Research Association.

Andrade, J. M. & Valentini, F. (2018). Diretrizes para construção de testes psicológicos: A Resolução CFP nº 009/2018 em destaque. *Psicologia: Ciência e Profissão, 38*(núm. esp.), 28-39. https://doi.org/10.1590/1982-3703000208890

Artino, A. R., Jr., Gehlbach, H., & Durning, S. J. (2011). AM Last Page: Avoiding five common pitfalls of survey design. Academic Medicine: *Journal Of The Association Of American Medical Colleges, 86*(10), 1327-1327. https://doi.org/10.1097/ACM.0b013e31822f77cc

Artino, A. R., Jr., & Gehlbach, H. (2012). AM last page: Avoiding four visual-design pitfalls in survey development. Academic Medicine: *Journal of the Association of American Medical Colleges, 8*7(10), 1452. https://doi.org/ 10.1097/ACM.0b013e31826ac042

Baka, A., & Figgou, L. (2012). "Neither agree, nor disagree": A critical analysis of the middle answer category in Voting Advice Applications. *International Journal of Eletronic Governance, 5*(3/4), 244-263. https://doi.org/10.1504/IJEG.2012.051306

Brown, A. (2014). Item Response Models for Forced-Choice Questionnaires: A Common Framework. *Psychometrika, 81*(1), 135-160. https://doi.org/10.1007/s11336-014-9434-9

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460-502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135-1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, Vol. 18,* pp. 36-52. https://doi.org/10.1037/a0030641

Brown, A., & Maydeu-Olivares, A. (2016). Modeling forced-choice response formats. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (pp. 1-64). London: John Wiley & Sons, Inc.

Cabooter, E., Weijters, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research, 69*(7), 2574-2584. https://doi.org/10.1016/j.jbusres.2015.10.138

Casper, W., Edwards, B. D., Wallace, J. C., Landis, R. S., & Fife, D. A. (2020). Selecting response anchors with equal intervals for summated rating scales. *Journal of Applied Psychology, 105*(4), 390. https://doi.org/10.1037/apl0000444

Cattell, R. B. (1958). What is "objective" in "objective personality tests?" *Journal of Counseling Psychology, 5*(4), 285-289. https://doi.org/10.1037/h0046268

Chan, D. (2009). So why ask me? Are self-report data really that bad? In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 309-336). New York, NY: Routledge/Taylor & Francis Group.

Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika, 30*(1), 99-121. https://doi.org/10.2333/bhmk.30.99

Chyung, S. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the Likert scale. *Performance Improvement, 56*(10), 15-23. https://doi.org/10.1002/pfi.21727

Conselho Federal de Psicologia. (2022). **Resolução n° 31, de 15 de dezembro de 2022.** Estabelece as diretrizes para a realização da Avaliação Psicológica no exercício profissional da psicóloga e do psicólogo, regulamenta o Sistema de Avaliação de Testes Psicológicos - SATEPSI e revoga a Resolução CFP n° 09/2018. Recuperado de https://atosoficiais.com.br/cfp/resolucao-do-exercicio-profissional-n-31-2022-estabelece-diretrizes-para-a-realizacao-de-avaliacao-psicologica-no-exercicio-profissional-da-psicologa-e-do-psicologo-regulamenta-o-sistema-de-avaliacao-de-testes-psicologicos-satepsi-e-revoga-a-resolucao-cfp-no-09-2018?origin=instituicao

Costa, A. R. L., & Hauck Filho, N. (2017). Menos desejabilidade social é mais desejável: Neutralização de instrumentos avaliativos de personalidade. *Interação em Psicologia, 21*(3). http://dx.doi.org/10.5380/psi.v21i3.53054

Dolnicar, S., & Grün, B. (2013). "Translating" between survey answer formats. *Journal of Business Research, 66*(9), 1298-1306. https://doi.org/10.1016/j.jbusres.2012.02.029

Furr, R. M. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology.* London: SAGE Publications.

Hauck-Filho, N., & Valentini, F. (2019). O controle da desejabilidade social no autorrelato usando quádruplas de itens. *Avaliação Psicológica, 18*(3), 1-3. https://dx.doi.org/10.15689/ap.2019.1803.ed

Hauck-Filho, N., Valentini, F., & Primi, R. (2021). Por que escalas balanceadas controlam a aquiescência nos escores brutos? *Avaliação Psicológica, 20*(1), a-c. https://dx.doi.org/10.15689/ap.2021.2001.ed

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, Vol. 91*, pp. 9-24. https://doi.org/10.1037/0021-9010.91.1.9

Hontangas, P. M., Leenen, I., Torre, J. De, Ponsoda, V., & Morillo, D. (2016). Traditional scores versus IRT estimates on forced-choice tests based. *Psicothema, 28*(1), 76-82. https://doi.org/10.7334/psicothema2015.204

Hontangas, P. M., Torre, J. De, Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement, 39*(8), 598-612. https://doi.org/10.1177/0146621615585851

Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment, 32*(7), 663-673. doi:10.1177/0734282914522200

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 140, 55. https://doi.org/2731047

Luft, J., & Ingham, H. (1961). The johari window. *Human Relations Training News, 5*(1), 6-7.

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology, 142*(2), 71-89. https://doi.org/10.1080/00221309.2014.994590

Parker, R. I., Vannest, K. J., & Davis, J. L. (2013). Reliability of multi-category rating scales. *Journal of School Psychology, 51*(2), 217-229. https://doi.org/10.1016/j.jsp.2012.12.003

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529-554.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). An introduction and a point of view. In R. Tourangeau, L. J. Rips, & K. Rasinski (Eds.), *The psychology of survey response* (pp. 1-22). Cambridge, UK: Cambridge University Press.

Valentini, F., & Hauck-Filho, N. (2020). O impacto da aquiescência na estimação de coeficientes de validade. *Avaliação Psicológica, 19*(1), 1-3. https://dx.doi.org/10.15689/ap.2020.1901.ed

Wang, W., Qiu, X., Chen, C., Ro, S., & Jin, K. (2017). Item response theory models for ipsative tests with Multidimensional Pairwise Comparison Items. *Applied Psychological Measurement, 41*(8), 1-14. https://doi.org/10.1177/0146621617703183

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236-247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Weijters, B., Geuens, M., & Baumgartner, H. (2013). The Effect of Familiarity with the Response Category Labels on Item Response to Likert Scales. *Journal of Consumer Research, 40*(2), 368-381. https://doi.org/10.1086/670394

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972. doi:10.1177/0013164404268674

Willits, F. K., Theodori, G. L., & Luloff, A. E. (2016). Another look at Likert scales. *Journal of Rural Social Sciences, 31*(3), 126-139. Retrieved from http://journalofruralsocialsciences.org/pages/Articles/JRSS 2016 31/3/JRSS 2016 31 3 126-139.pdf

recebido em novembro de 2020
aprovado em agosto de 2021

Sobre os autores

Ariela Raissa Lima-Costa é Psicóloga, Mestre e Doutora em Psicologia pela Universidade São Francisco. Atualmente é Professora adjunta no Programa de Pós-Graduação Psicologia da Universidade São Francisco.

Bruno Bonfá-Araujo é Psicólogo, Mestre e Doutor em Psicologia pela Universidade São Francisco. Atualmente é Pós-Doutor da University of Western Ontario.

Como citar este artigo

Lima-Costa, A. R., & Bonfá-Araujo, B. (2022). Construindo Escalas de Autorrelato: O que fazer? *Avaliação Psicológica, 21*(3), 329-338. http://dx.doi.org/10.15689/ap.2022.2103.21860.09