Análise Psicológica (2020), 2 (XXXVIII): 271-288 doi: 10.14417/ap.1659

O formato das questões de resposta fechada: Implicações para a natureza, validade e fiabilidade das medidas

Teresa Garcia-Marques[*] / Rui Bártolo-Ribeiro[**]

[*] William James Center for Research, ISPA – Instituto Universitário, Lisboa, Portugal; [**] APPsyCI – Applied Psychology Research Center Capabilities & Inclusion, ISPA – Instituto Universitário, Lisboa, Portugal

Este artigo tem por objectivo apoiar os investigadores nas decisões sobre o uso de questões de resposta fechada nos seus questionários fazendo uma revisão crítica da literatura relativamente às implicações dessas decisões para a natureza, validade e fiabilidade da medida e das conclusões que ela suporte. Esta revisão da literatura apoia a investigador no processo de decisão sobre a melhor forma de operacionalizar as suas variáveis através de uma resposta fechada. São apresentados argumentos que sustentam a tomada de decisão na construção da lista de opções de resposta a fornecer ao inquirido, e o tipo de escala a utilizar: gráficas ou não gráficas; categorias ou de avaliação contínua; com 3 ou mais pontos; com ou sem rótulos e neste caso, com que tipo de rótulos, etc. Ilustramos adicionalmente outros tópicos a ter em conta na construção de uma medida que usa como formato de resposta fechado uma escala de avaliação contínua analisando o caso específico em que se mede "frequências percebidas".

Palavras-chave: Escalas de avaliação, Questionários, Enviesamento.

É-nos já muito familiar o responder a um questionário situando a nossa resposta entre dois polos de um contínuo. Por exemplo, a uma questão sobre o grau de interesse que o título deste artigo lhe suscitou poderia pedir-lhe uma resposta situada numa escala que vai desde 1 – *Nada interessante* a 7 – *Muito interessante*, fazendo um círculo em torno do número, uma cruz no quadrado, ou um *click* no círculo correspondente ao número que melhor representasse a resposta. Não nos espantaríamos, porém se associado ao número, ao quadrado ou ao círculo que representasse o centro do contínuo (e.g., 1 a 7) surgisse o rótulo *Nem interessante, Nem desinteressante*. Na realidade poderíamos até encontrar esse tipo de rótulos em todos os pontos do contínuo. Porém alguns de nós poderemos considerar mais fácil informar sobre o grau de interesse suscitado pelo título do artigo através da manifestação de acordo com uma afirmação do tipo *"Acho o assunto abordado muito interessante"*, seleccionando uma resposta numa escala ancorada em 1 – *Discordo totalmente* a 7 – *Concordo totalmente*[1].

William James Center for Research, ISPA – Instituto Universitário, é financiado pela Fundação para a Ciência e Tecnologia (Ref. UID/PSI/04810/2013).

A correspondência relativa a este artigo deverá ser enviada para: Rui Bártolo-Ribeiro, APPsyCI – Applied Psychology Research Center Capabilities & Inclusion, ISPA – Instituto Universitário, Rua Jardim do Tabaco, 34, 1149-041 Lisboa, Portugal. E-mail: rbartolo@ispa.pt

[1] O facto de Likert ter sugerido o uso deste tipo de itens associados a uma escala de avaliação de concordância ("rating scales") no procedimento que referiu ser necessário à composição de uma "escala de Likert" (escalas que envolve vários itens para aceder a um mesmo construto, sendo estes itens avaliados em escalas de 5-7 pontos e cuja média das avaliações é pressuposta representar o constructo a ser medido) leva alguns autores a designar estas *rating scales* como escalas *tipo Likert*.

Quem constrói um questionário tem de decidir sobre o formato deste tipo de questão (questão de resposta fechada, ver abaixo) e para tal deverá ter em conta os objectivos do inquérito as variáveis que pretende medir e garantir de que vai construir uma boa medida (representativa, única, válida e fiável) dessas variáveis e não seleccionar um ou outro tipo de formato de resposta por questões estéticas e de facilidade de concretização. A preocupação estética advém da necessidade que os investigadores sentem de envolver os inquiridos nas suas respostas, embora possa haver consequências associadas a estas opções.

Neste artigo chamamos a atenção do investigador para as consequências das suas decisões, com base no que nos é sugerido pela investigação sobre inquéritos.

Comecemos por esclarecer o conceito de "medida", questões sobre a sua validade e fiabilidade e sobre como a sua natureza se relaciona com a análise estatística. Apresentamos de seguida os principais tipos de respostas fechada (na sua versão normal e versão gráfica), referindo as propriedades das medidas que geram.

Medida e suas propriedades

A medição de uma variável define o processo de atribuição de um valor (e.g., um número) cuja variação estabelece uma relação directa com a própria variável. Idealmente (princípio de *representação*) procura-se estabelecer uma relação "ponto por ponto" entre a variável e os números, para que as relações entre os números reflictam as relações entre as instâncias medidas dessa variável; e que essa relação não seja perdida, mantendo-se única, sob qualquer transformação realizada sobre esses números (princípio da *unicidade*) (ver Hand, 1996, para outras concepções de "medida"; Michell, 1986). O processo de atribuição destes números deverá garantir que a relação qualitativa entre variáveis possa ser verificada na relação que se estabelece entre os números e assuma as propriedades do sistema numérico (Townsend & Ashby, 1984).

As duas características fundamentais de uma medida são a sua *validade* e a sua *fiabilidade*. Uma medida é tanto mais válida quanto maior a certeza de que ela é uma representação do atributo em causa e apenas desse. Por outro lado, uma medida é tanto mais fiável quanto maior confiança houver de que o número obtido num dado momento será replicado noutras medições do mesmo atributo se se mantiverem as mesmas condições. Os investigadores em psicometria desenvolveram várias formas de aceder a estas duas características da medida, nomeando-as de diferentes "validades" (e.g., validade de conteúdo, de constructo, relativa ao critério) e diferentes "fiabilidades" (e.g., consistência interna, estabilidade temporal) com recurso a diversos índices estatísticos para nos informar dessas propriedades. A maioria das abordagens psicométricas operacionalizam uma medida através de "escalas de múltiplos itens" (essencialmente as escalas construídas segundo a metodologia proposta por Likert), qualquer questão fechada deve ser caracterizada na sua validade e fiabilidade. Tomemos, por exemplo, a simples questão sobre o género do inquirido. Um investigador ao dar como opções de resposta "Feminino" e "Masculino", pretendendo conhecer o género de nascença do inquirido, pode confrontar-se com a falta de validade da medida, por algumas das respostas obtidas reflectirem o "género psicológico", em vez do pretendido "género de nascença" do inquirido.

O uso de questões de resposta fechada

Ao se colocar mais de uma questão a uma pessoa, realizamos um inquérito que se for apresentado oralmente é designado por *entrevista* e por escrito é designado de *questionário*. A estrutura das questões de um questionário tende a ser mais rígida do que na entrevista, dado que se define *a priori*

a ordem das questões, a dimensão e o tipo de respostas. Quando se limita o tipo de resposta a dar pelo inquirido, referimos a questão como de resposta "fechada". Contrariamente às questões de resposta aberta que permitem ao inquirido responder com as suas próprias palavras, nas questões de resposta fechada o inquirido selecciona a opção (de entre as apresentadas), que melhor representa a sua opinião.

As respostas abertas têm maior relevância em investigação qualitativa (e.g., *grounded theory*) e exploratória. Na realidade, quando *a priori* há pouco conhecimento sobre o constructo em estudo deve-se aceder ao maior número possível de unidades informativas numa só resposta (a ser sujeita a uma análise de conteúdo). Essas unidades informativas definem as variáveis que serão analisadas. Futuramente poder-se-á construir uma medida dessas variáveis que podem assumir uma natureza complexa de múltiplos itens (e.g., escalas de Likert, diferenciais semânticos) ou serem definidas em simples questões de resposta fechada. Quando na investigação já se sabe *a priori* que unidade(s) informativa(s) sustentam as variáveis a medir, a opção por colocar uma questão (ou mais) de resposta fechada pode ser vantajosa. Um aspecto ligado à validade desta medida (como salientamos ao longo deste artigo), é que é agora o próprio inquirido quem classifica o conteúdo das suas respostas, facilitando a tarefa, tanto do inquirido como do investigador.

Uma questão de resposta fechada é (ou contribui para) a operacionalização da variável em estudo, definindo as características dessa medida e por tal as conclusões a serem retiradas. Tomemos, como exemplo, um estudo sobre "hábitos de leitura", que nos questiona sobre a frequência com que lemos romances, policiais, revistas, jornais etc. Apesar da questão informar sobre os hábitos de leitura, ela não permite concluir, por exemplo, sobre quanto lemos no nosso dia-a-dia. Para esse objectivo a medida deveria focar todas as fontes de leitura, desde os menus dos restaurantes às legendas dos filmes e internet. Também não informa sobre a importância dada à leitura, dado que a frequência de leitura, não traduz obrigatoriamente a importância. Em função do objectivo do estudo, define-se a variável em estudo. A sua operacionalização através de uma questão de resposta fechada deve procurar ser única (focada apenas num atributo), exaustiva e ter validade facial (questão adequada ao atributo medido). Após definir a questão há que determinar o tipo de resposta pretendido. Uma resposta por selecção de uma opção de entre uma lista, ou posicionar a opção de resposta numa ordem ou num contínuo. Analisamos de seguida cada um destes tipos de resposta fechada, seus formatos e referimos os problemas que levantam à validade das conclusões que delas retiramos. A literatura não é nem sistemática nem exaustiva nesta análise, mas sempre que possível fornecemos dados que demonstram a implicação de uma dimensão do formato de resposta para a validade e fiabilidade da medida e das conclusões que estas sustentam.

Resposta por selecção de entre uma lista de alternativas

Uma lista de opções de resposta deve ser exaustiva e ter as regras bem definidas a que deve estar sujeita a selecção: se apenas uma ou se mais opções devem ser seleccionadas e/ou se estas devem ser ordenadas. A selecção de uma opção entre várias define a medida como tendo uma natureza *nominal*. Este é tipicamente o caso da operacionalização de variáveis como o género, o estado civil, a definição da posição que ocupa numa empresa, ou do tipo de emoção que melhor representa o que sente num dado momento. Nestas apresenta-se como opções de resposta, ser homem *versus* mulher; casado, solteiro, viúvo ou divorciado; chefia ou subordinado; e tristeza, alegria, raiva etc. Quando é possível seleccionar mais de uma opção, cada uma das opções listadas define-se como uma variável dicotómica; foi *versus* não foi seleccionada (valor 1 ou 0). Assim, se no estudo sobre hábitos de leitura seleccionássemos o romance e não o jornal informávamos o investigador que lemos o primeiro e não o segundo. Contudo, se a regra de resposta incluísse a ordenação das diferentes opções, perceberíamos, por exemplo que o inquirido lê mais romances do que policiais. Neste caso, o valor atribuído a cada variável é ordinal. Importante e limitativo

das opções de análise, a informação fornecida é totalmente dependente da lista das opções, não sendo as medidas independentes (e.g., *No menu fornecido eu coloquei o prato de peixe em segundo lugar, mas apenas porque me ofereceram a possibilidade de comer lagosta*).

Vários estudos demonstram que tanto a lista de opções, como a ordem da sua apresentação, condicionam as respostas dos inquiridos comprometendo a validade das conclusões dos estudos (ver Schwarz & Hippler, 1990). Assim, a importância da exaustividade para a validade de conteúdo, dado que mesmo adicionando como alternativa de resposta a categoria "Outras", o inquirido tende a utilizar apenas as respostas disponibilizadas e a procurar entre elas a opção que melhor se adapta ao seu caso (Lindzey & Guest, 1951; Schuman & Scott, 1987). A opção de resposta "Outras", como oportunidade de designação, tem maior relevo em pré-testes, onde se pode explicitamente pedir aos inquiridos para completar a lista de opções oferecidas, e assim inclui-las numa lista mais exaustiva no estudo.

O efeito da ordem em que as opções de resposta são apresentadas é usualmente um efeito de primazia (os primeiros itens têm maior probabilidade de serem seleccionados por receberem maior atenção, ver Schwarz & Hippler, 1990). Para anular ou isolar (se necessário) o efeito, a ordem deve ser aleatória ou contrabalançada (com custos no número de participantes e na análise estatística).

A lista de opções de resposta cria um contexto com uma influência directa nas escolhas dos respondentes. O exemplo mais típico deste tipo de efeitos é o efeito de "contraste". Os estudos de Noelle-Neumann (1970) ilustram este efeito. Eles inquiriram sobre que itens de uma lista seriam "tipicamente alemães". Os inquiridos consideraram como menos típicos o "macarrão" e as "batatas", quando estes foram precedidos da opção "arroz" (um alimento muito usado pelos alemães). Outro exemplo, é o estudo de Schwarz e Hippler (1990) que demonstra que se um item extremamente positivo/negativo é apresentado na lista, a probabilidade de itens moderados subsequentes serem seleccionados como positivos/negativos diminui. Estes factores influenciadores das respostas devem ser considerados, quer no momento de construção do questionário, quer na interpretação dos resultados.

Respostas por posicionamento numa ordem ou num contínuo (rating scales)

A medida de um atributo que varia em termos ordinais ou num contínuo deverá ser uma questão que solicita uma resposta que varia da mesma forma. A Figura 1 apresenta alguns exemplos dos diferentes tipos de escalas de categorias ordinais e escalas de avaliação contínua, bem como algumas versões gráficas (e.g., respostas num contínuo unidimensional medidas em centímetros, a *Visual Analogue Scale – VAS*). Analisamos de seguida as suas propriedades métricas (validade e fiabilidade).

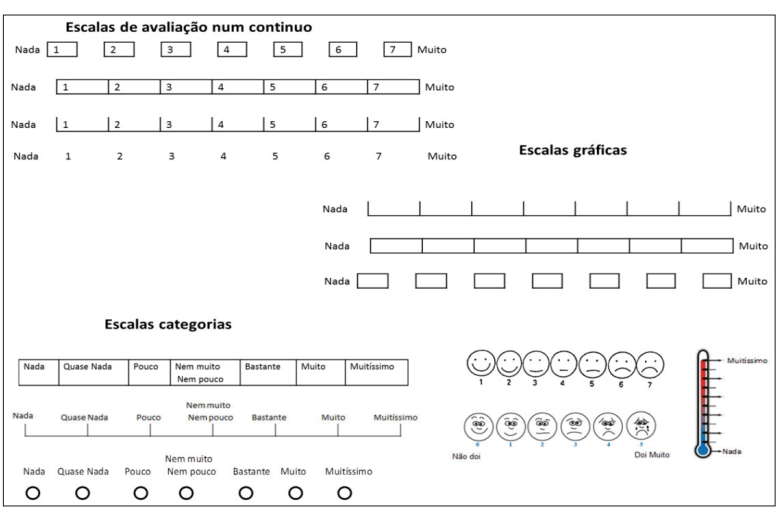

Figura 1. Escalas de avaliação num contínuo, escalas de categorias e escalas gráficas

Qualquer das escalas representadas na Figura 1, são "medidas truncadas" da dimensão psicológica medida. O contínuo onde se definir a medida, situa-se entre dois valores que podem ser ou não ser extremos. O nível de truncamento é definido pelos rótulos/âncoras fornecidos nos pontos extremos. A escala pode ir de *Pouco* a *Muito* ou de *Nada* a *Muitíssimo*. Sendo o mesmo o número de pontos da escala, a primeira tem mais sensibilidade a medir respostas de intensidade mediana (visto os números representarem maior variabilidade em valores médios) e menos a medir os extremos (onde várias intensidades serão representadas pelo mesmo número; o *Nada* e *Pouco* serão representados pelo valor 1, assim como *Muito* e *Muitíssimo* pelo valor mais elevado). Outro problema que advém do facto das escalas serem truncadas é a possível criação de uma assimetria na distribuição das respostas, pondo em causa a sua aproximação à curva normal. Um pré-teste à escala pode esclarecer a necessidade de se truncar de forma diferente a medida. Por exemplo, ao perceber que todos os inquiridos têm uma atitude positiva, pode-se optar por usar uma escala que acede apenas à dimensão positiva, mas com maior sensibilidade (e.g., 1 – *Gosto Pouco* a 7 – *Gosto muitíssimo*, em vez de 1 – *Não gosto* a 7 – *Gosto*). O facto de as escalas serem truncadas não invalida por si só, as comparações entre respostas de inquiridos; mas respostas em escalas truncadas de forma diferente não devem ser levianamente comparadas.

Qualquer escala da Figura 1 informa sobre a grandeza ou intensidade da variável. Todas podem medir, por exemplo, a *intensidade da emoção*. Para isso as suas âncoras terão de representar apenas e só a dimensão "intensidade" (e.g., "baixo" – "elevado"). Mas as escalas diferem na forma como representam a natureza intervalar dessa *intensidade*. As respostas numa escala de avaliação contínua reflectem melhor a "igualdade de intervalos" entre níveis de intensidade do que as respostas em categorias ordinais. Neste sentido, são uma "melhor operacionalização" de uma variável que pressupomos ser contínua.

Vários factores contribuem para a validade, fiabilidade e a natureza das medidas que usam as escalas da Figura 1. Abaixo sumariamos os dados actualmente disponíveis na literatura, que exemplificam o papel relevante destes formatos.

Uso do conhecimento prévio sobre continuidade

O modo como os inquiridos percebem cada escala influencia a natureza mais ou menos intervalar das respostas obtidas através dessa escala. É porque todos temos a noção de número e de espaço que a percepção de uma continuidade com igualdade de intervalos ocorre especialmente em escalas que usam números a representar cada nível de medida, ou que fazem uso do "espaço geométrico" (e.g., a VAS; ver Hayes & Paterson, 1921) para representar a continuidade. Nelas o inquirido percebe e mantém a ideia de equidistância entre os pontos, garantindo que as suas respostas definam um contínuo (ver Kiess & Bloomquist, 1985). Porque aprendemos a natureza contínua dos números e do espaço (sabemos que 2 é o dobro de 1 e que 8 é o dobro de 4, etc.), quando nos pedem para nos situarmos nesse espaço aplicamos esse conhecimento.

É porque se sabe que usamos os conhecimentos prévios para responder a estas escalas que se têm desenvolvido as escalas gráficas (ver Converse & Presser, 1986) e que elas tendem a ser boas representações da continuidade. As escalas gráficas usam contínuos existentes no mundo para nos ajudar a representar a nossa resposta, sendo o caso mais clássico o uso do termómetro que usualmente mede a temperatura, e foi usado para medir atitudes (e.g., Berman & Stookey, 1980). Estudos sobre as propriedades da VAS, sugerem que estas escalas garantem a linearidade da medida (Hofmans & Theuns, 2008; Myles, Troedel, Boquest, & Reeves, 1999; Myles & Urquhart, 2005), que mantém a igualdade de intervalos (Reips & Funke, 2008) e que são escalas sensíveis (Abend, Dan, Maoz, Raz, & Bar-Haim, 2014; Rausch & Zehetleitner, 2014). Ao rever de forma sistemática as propriedades métricas de escalas VAS de várias medidas (e.g., estado de espírito; dor), Wewers e Lowe (1990), concluem que elas fornecem medidas válidas (validade de

constructo, validade de critério e validade discriminante) e fiáveis (teste-reteste) desses construtos (mas ver McCormack, Horne, & Sheather, 1988, para uma posição contrária).

Existem igualmente evidências de que que a VAS se comporta de forma semelhante a escalas contínuas como no caso dos "botões" (ver Couper, Tourangeau, Conrad, & Singer, 2006; Funke & Reips, 2012). Também os estudos que comparam as escalas numéricas com a VAS sugerem a sua equivalência (e.g., Averbuch & Katzper, 2004; Cork et al., 2004; Funke & Reips, 2012; Flynn, van Schaik, & van Wersh, 2004), embora pareça existir uma superioridade das escalas numéricas, na garantia da igualdade de intervalos e numa melhor aproximação a uma distribuição normal (e.g., Paul-Dauphin, Guillemin, Virion, & Briançon, 1999). Porém, o bom desempenho das VAS não é reflectido nas suas versões dinâmicas digitais, os *sliders.* Comparativamente à VAS, os *sliders* fornecem um desempenho muito inferior, pelo que segundo Funke (2016) é de todo recomendado evitar o seu uso.

Falta de ponto de referência (o zero)

Mesmo as escalas que são passíveis de ter intervalos constantes entre os números não nos garantem que essa grandeza é representada de forma igual por todos os indivíduos. Nenhuma das escalas por defeito, operacionaliza uma medida com "um zero" (a ausência do constructo). Como consequência, cada indivíduo poderá perceber de forma diferente, por exemplo, o que é uma "baixa intensidade emocional" calibrando idiossincraticamente as suas respostas. Tal limita as conclusões que podemos retirar dos dados. Por exemplo, ao pedirmos a um trabalhador para, numa escala que varia de 1 – *Muito Pouco* a 7 – *Muitissimo,* assinalar a "quantidade de trabalho" que realizou naquele dia*,* o trabalhador A (que arquivou 10 ficheiros) pode reportar o valor 4 na escala, e o trabalhador B (que arquivou 20 ficheiros) reportar o valor 2. As respostas reflectem a forma como cada trabalhador se situa num espaço psicológico subjectivo; um espaço do que é para cada um deles "pouco" ou "muito". E deverá ser assim que o investigador interpreta os seus dados, não confundido o espaço criado subjectivamente com o espaço objectivo, que exige a utilização de uma escala de razão com clara definição do que é um zero absoluto. Para isso, ajuda (mas não chega) fornecer aos inquiridos a operacionalização dos intervalos, por exemplo, referir que "Muito pouco=<5 ficheiros".

Uso de rótulos verbais

Se os inquiridos atenderem aos rótulos verbais (ou quantificadores, designações), que um investigador decide usar para referenciar cada ponto da escala (que não só os extremos), irão fornecer respostas diferentes do que se não atenderem. Tal introduz um erro sistemático na medida, dado que nunca saberemos o que o inquirido fez (ver Cummins & Gullone, 2000).

O uso de rótulos também é uma ameaça à continuidade da medida. Quer os quantificadores, quer os rótulos que os representam, garantem a ordenação dos pontos de mensuração dando-lhes um significado (validade "facial" de uma resposta), mas perturbam a equidistância dos intervalos. A adição de rótulos perturba o nível de mensuração da medida conferindo-lhe um estatuto de escala ordinal[2] (e.g., Hartley, Trueman, & Rodgers, 1984; Meek, Sennot-Miller, & Ferketich, 1992; Rasmussen, 1989; Wills & Moore, 1994).

[2] Aqui referimos o nível de mensuração de uma "escala de resposta". Esta afirmação não deve ser confundida com o nível de mensuração que tem a média de diferentes respostas (idealmente de 20-30 itens, como defendido por Likert). O teorema de limite central garante-nos que a distribuição e amostragem das médias desses itens terão uma distribuição normalizada e por tal as médias em si mantêm distâncias intervalares.

A natureza dos rótulos verbais utilizados

Todo o rótulo atribuído a um ponto da escala modela o significado atribuído a esse ponto. Os rótulos dos extremos numa escala contínua definem a unicidade da medida (ver Cummins & Gullone, 2000) e deve ser sujeita a certificação em pré-testes (garantido que todos os inquiridos os interpretam da mesma forma). Por exemplo, para garantir que os diferenciais semânticos definem uma escala contínua, com uma única dimensão semântica, Osgood, Suci e Tannenbaum (1957) usaram como âncoras, adjectivos de polos opostos (Bom *vs.* Mau; Grande *vs.* Pequeno; Forte *vs.* Fraco etc.). A vantagem dos rótulos-âncora serem verdadeiros opostos semânticos, e serem *percebidos* como opostos é que garantem a simetria da medida (Worcester & Burns, 1975). Mas tal nem sempre é claro. Por exemplo, o que é o oposto de Nada? Algo? Muito? Muitíssimo?

Também, a escolha dos números onde ancoraram a escala influencia as respostas dos inquiridos. Por exemplo, num estudo clássico nesta área, Schwarz, Knauper, Hippler, Noelle-Neuman e Clark (1991) verificaram que enquanto 34% dos inquiridos referiam ser "bem-sucedidos" na sua vida, numa escala ancorada entre -5 (*Nada bem-sucedido*) e 5 (*Extremamente bem-sucedido*), apenas 3% reportaram esse nível de sucesso numa escala ancorada em 0 (*Nada bem-sucedido*) e 10 (*Extremamente bem-sucedido*). Pelo que parece, o significado de *Nada bem-sucedido* quando associado ao número zero, refere a ausência de sucessos e quando associado ao número menos cinco (-5) refere a presença de fracassos. Para que um investigador saiba como interpretar os seus dados tem de perceber o significado que os seus respondentes atribuem às suas respostas, o que é apenas identificado por um processo cuidadoso de pré-testes.

A dimensão mais utilizada em escalas é o grau de acordo manifestado numa afirmação, operacionalizado frequentemente em escalas de categorias do tipo das usadas por Likert (1932) na construção dos seus instrumentos de medida (5 pontos que associavam cada ponto a um rótulo: *1 – Discordo totalmente, 2 – Discordo, 3 – Nem discordo nem concordo, 4 – Concordo, 5 – Concordo totalmente*). Mas o uso de diferentes rótulos pode alterar as respostas sem que tal reflicta diferenças na posição do inquirido. Por exemplo, se em vez do rótulo "*Discordo"* o investigador usar o rótulo "*Discordo bastante"*, irá deslocar a resposta de alguns inquiridos para os pontos adjacentes e vice-versa.

Sustentados em Likert, muitos investigadores usam escalas de categorias adicionando diferentes tipos de rótulos aos seus pontos. Entendem que tal facto reduz a ambiguidade das respostas dos inquiridos e confere maior estabilidade à medida (Svensson, 2000). Mas a verdade é que os problemas com o seu uso, limitam essas aparentes vantagens. Facilmente se identificam dois problemas: (a) os rótulos quebram a segurança de que os intervalos são percebidos como iguais e (b) os rótulos violam a presunção de invariabilidade e estabilidade do significado atribuído a cada ponto. Ilustremos os dois problemas tomando como exemplo a dimensão favorabilidade e o uso dos seguintes quantificadores linguísticos: *1 – Totalmente favorável; Bastante favorável; Moderadamente favorável; Nem favorável nem desfavorável; Moderadamente desfavorável; Bastante desfavorável; 7 – Totalmente desfavorável.* Esta escala pressupõe que a distância psicológica percebida entre "totalmente" e "bastante" seja idêntica a entre "bastante" e "moderadamente" e presume que todos os participantes percebam o ponto 6 como superior ao ponto 5 (ou seja, que ninguém entenda a palavra *bastante* como de valor superior a *moderadamente*). Os estudos (e.g., Schriesheim & Novelli, 1989) têm porém demonstrado que não existe consenso entre o que representa uma grandeza maior, o "bastante" ou o "moderadamente". E o advérbio de tempo "ocasionalmente", foi identificado como muito diferente do advérbio "raramente", mas relativamente próximo, em significado, ao "às vezes". O mapeamento gráfico da própria escala ajuda a eliminar estes problemas. Assim, o contínuo com igualdade de intervalos, apenas é garantido se os inquiridos ignorarem o significado semântico dos rótulos. Mas ao colocar os rótulos, uns inquiridos podem ignorá-los e outros não, inserindo uma fonte de erro na medida.

Vários estudos têm procurado estudar os quantificadores ou rótulos a serem usados nas escalas de categorias com vista a ultrapassar a natureza discreta da variável (e.g., Osinski & Bruno, 1998; Schriesheim & Castro, 1996; Schriesheim & Schriesheim, 1974). Alguns destes estudos seguem a orientação de comparação diáticas sugeridas por Thurstone (ver, por exemplo, Myers & Warner, 1968; Wildt & Mazis, 1978), enquanto outros seguem a sugestão de Stevens (1975) e apresentam aos participantes simultaneamente os diferentes quantificadores de resposta, pedindo-lhes para atribuir um valor a cada, e analisando-se de seguida a natureza de relação estabelecida, procurando perceber a invariância e as distâncias que se estabelecem entre eles. Um exemplo é o estudo de Bartram e Yelding (1973), que testou expressões como *Extremamente, Muito, Bem*, *Normalmente*, *Razoavelmente*, *Quase*, e *Não em todos* com vista a perceber se mantinham a invariância da medida e uma equidistância percebida. Mais tarde Schriesheim e Novelli (1989) usaram os dois métodos para o mesmo conjunto de rótulos e concluíram que estes fornecem diferentes conclusões. A lição que se tira destes estudos é que os rótulos mudam as distâncias entre os números. Por exemplo, o estudo de Ware e Gander (1994) usando o método Thurstone sugerem as seguintes distâncias entre os rótulos de uma escala categorial: *Poor (1.0), Fair (2.3), Good (3.4), Very Good (4.3), and Excellent (5.0),* o que deixa claro que a distância entre as duas categorias mais baixas (1.3) é aproximadamente o dobro entre as duas categorias mais altas (0.7).

Para o leitor português é talvez mais relevante conhecer os resultados dos trabalhos de Osinski e Bruno (1998) que replicaram os trabalhos de Schriesheim e Novelli (1989). Os autores estudaram 20 expressões de frequência, tais como: *Sempre, Constantemente, Continuamente, Frequentemente, Raramente, Não em todos, Nenhuma das vezes, Nunca*, e sugerem como podemos transformar as escalas de categorias em escalas com pontos não equidistantes, mas bem definidos num *continuum* (como sugerido por Thurstone, 1928). Por exemplo, para uma escala de 5 pontos, os autores recomendam o uso de rótulos em que o participante que seleccionar a opção de resposta *Quase nunca* terá uma pontuação de 2.28; *Às vezes* uma pontuação de 4.71; *Normalmen*te uma pontuação de 26.60; *Quase sempre* uma pontuação de 38.49; e *Sempre* uma pontuação de 50.38.

O número de pontos definidos na escala de resposta

Não nos surpreende uma escala de resposta ter 5, 7, ou ainda 11 pontos. Estes são valores tradicionais, associados a autores como Likert (5 pontos), Osgood, Suci e Tannenbaum's (7 pontos) e Thurstone (11 pontos). A analogia da escala de resposta com um termómetro, introduziu a escala como 10 ou 100 pontos.

Uma escala de resposta com mais pontos tem maior sensibilidade pelo que por defeito poderia enriquecer as análises de dados, facilitando as análises multivariadas (ver Viswanathan, Sudman, & Johnson, 2004). Porém, um aumento de sensibilidade nem sempre é desejável pois pode predispor a medida a erro se a nossa actividade cognitiva não acompanha essa mesma sensibilidade. A consciência deste facto levou a que muitos estudos tenham salientado a alteração da fiabilidade das respostas em escalas de avaliação contínua com a alteração de número de pontos. Os primeiros estudos apontaram para uma solução em torno de 5-9 pontos (e.g., Andrews & Withey 1976; Cox 1980; Givon & Shapira 1984; Jenkins & Taber, 1977; Neuman & Neuman 1981).

Em 1984, Churchill e Peter conduziram uma meta-análise onde verificaram que a fiabilidade de uma medida compósita de itens, que envolve o somatório de diferentes escalas de resposta (associados a diferentes itens) aumenta directamente com o número de pontos da escala de resposta. Contudo, usando o método de simulação Monte-Carlo, Cicchetti, Showalter e Tyrer (1985) contradizem essa conclusão sugerindo que essa relação linear é verificada apenas com escalas de resposta entre dois e sete pontos. Entre sete e 100 categorias de respostas não se verifica qualquer incremento de fiabilidade (dados replicados por Preston & Colman, 2000). A escala de sete pontos ficou assim destacada na sua fiabilidade. Autores como Srinivasan e Basu (1989) e Oaster (1989) que testaram a fiabilidade quer por meio do teste-reteste quer pela medida de consistência interna

(alfa de Cronbach) confirmam tal facto. Estudos mais recentes apontam para valores desejáveis de fiabilidade nas escalas que têm de 5 a 9 pontos (ver Alwin & Krosnick, 1991; Cicchetti et al., 1985; Colman, Norris, & Preston, 1997; Preston & Colman, 2000).

Com o objectivo de perceber qual é o número de pontos que reflecte a resposta natural dos participantes, alguns estudos usaram as VAS pedindo aos inquiridos para colocarem livremente as suas respostas em linhas de 9 a 10cm. Estudaram de seguida qual seria a fiabilidade da medida se organizassem as respostas obtidas em 2 ou mais *cluster*s de pontos. No estudo pioneiro nesta metodologia, Champney e Marshall (1939), mostraram que a fiabilidade cresce dramaticamente quando se sobe de dois pontos para a 9. Verificou-se ainda alguma alteração ligeira até 18 pontos, mas nenhuma vantagem em aumentar o número de *clusters* a partir desse ponto. McKelvie (1978) ao se preocupar em saber qual o número de pontos que os inqueridos espontaneamente usam nessa linha, concluiu que as suas respostas se organizam em torno de 5 *clusters*, indicando os 5 pontos como uma "resposta natural" e por tal mais fácil para os inquiridos.

É importante perceber que as diferenças de fiabilidade das resposta em formato de escalas com diferente tamanho reflectem tanto as diferenças na sensibilidade, como na facilidade com que os inquiridos respondem. Isto porque qualquer factor que crie dificuldades aos inquiridos para fornecer a sua resposta, irá afectar a fiabilidade da medida. Foi assim, que quando Alwin (1997) ao comparar o desempenho de uma escala de avaliação contínua com 7 pontos com o desempenho de uma escala gráfica de termómetro, concluiu que junto da população geral esta última origina respostas mais fiáveis. Parece que para a população em geral é mais fácil usar um termómetro do que uma escala contínua.

Mas a facilidade de resposta num contínuo depende também do julgamento específico em causa. Se é fácil dizer se gostamos ou não gostamos do nosso emprego, torna-se mais difícil dizer quanto gostamos, tentando diferenciar entre o que significa gostar 8 ou gostar 7 numa escala de 20 pontos. A tarefa cognitiva tona-se exigente e o erro provável. Quando o número de opções é muito grande, tem-se verificado que o inquirido desenvolve uma pré-disposição para manter a mesma resposta ao longo do instrumento (Swait & Adarnowicz, 2001; Weathers, Sharma, & Niedrich, 2005). No entanto, o inverso pode ser verificado, e sentirmos dificuldade em comunicar quanto gostamos do emprego quando o valor 5 é o máximo e não queríamos assinalar o 4 por ser apenas um valor acima do ponto médio. O facto de a escala ser pouco sensível e não detectar pequenas diferenças, dificulta a resposta de quem está motivado para fazer análises mais finas. Deste exemplo, fica claro que para além de devermos atender à capacidade cognitiva do inquirido para decidir o número óptimo de pontos de uma escala, devemos também considerar o seu grau de motivação (ver Alwin, 1991; Krosnick & Alwin, 1989; Tourangeau, 1984).

Tendo em conta todos os factores que têm sido apontados, é mais ou menos consensual que relativamente a uma questão fechada o uso de escalas de respostas com 5 e 7 pontos têm vantagens relativamente a escalas contínuas com mais ou menos pontos (ver Cummins & Gullone, 2000).

Um investigador deverá sempre ter em conta que existem vantagens em manter o número de pontos das escalas num mesmo estudo, constante. A relação entre escalas com diferente número de pontos tende a não ser linear, tornado problemático o uso de correlações lineares. Como demonstra Kennedy, Riquier e Sharp (1996), a relação entre escalas de tamanho diferente tende a expressar-se numa relação curvilínea (*U-shaped*). Os seus resultados mostram que ao relacionar dados de uma escala de 10 pontos com os de uma de 5 pontos, a relação é mais forte nos números acima do valor médio de cada escala, tendendo a ser ligeiramente negativa ou nula para os números abaixo do ponto médio. O facto de a relação positiva ser mais forte que a relação neutra/negativa pode sugerir que as escalas de 5 pontos são menos equilibradas (o ponto médio é percebido como deslocado para a esquerda).

O tamanho da linha que representa a VAS é análogo ao número de pontos de uma escala. Linhas com menos de 10 cm tendem a introduzir maior variabilidade nas respostas, o que reflecte variância aleatória, elevando a probabilidade de erro da medida (e.g., Revill, Robinson, Rosen, &

Hogg, 1976). Porque a diferença de erros de medida em linhas horizontais de 10, 15 e 20 cm, são insignificantes, as de 10 cm de comprimento têm sido privilegiadas. É relevante tornar clara a grandeza da linha, pelo que as âncoras descritivas colocadas na continuação da linha, devem deixar um espaço para tornar claro onde esta termina (Huskisson, 1974). É pedido ao inquirido que faça a sua avaliação marcando um ponto na linha, colocando um x ou um traço sobre a linha.

O ponto médio das escalas

A maioria dos investigadores propuseram o uso de escalas de resposta com um número ímpar de pontos, definindo claramente um ponto médio. A principal vantagem em se usar este ponto médio (ver Cummins & Gullone 2000; Krosnick & Presser, 2010) é a de que o inquirido não é forçado na escolha, sentindo-se mais confortáveis em responder. Tal fica mais claro quando ao ponto médio é adicionado por exemplo o rótulo *Nem satisfeito, nem insatisfeito* numa escala que vai de *Muito insatisfeito* a *Muito satisfeito.*

A revisão de literatura dos estudos nesta área realizada por Chyung, Roberts, Swanson e Hankinson (2017) define três aspectos a serem levados em consideração quando se toma a decisão de se usar ou não uma escala de resposta ímpar: (a) a não inclusão do ponto médio põe em causa a igualdade de intervalos; (b) quando o ponto médio é considerado, ele tenderá a ser usado; e (c) as respostas no ponto médio nem sempre reflectem uma opinião por parte do inquirido. Tendem a ser um refúgio para manifestar a não vontade de responder ou uma estratégia de resposta que não requer reflecção. Chyung et al. (2017) sugerem que, por causa desta tendência de resposta, se omita o ponto médio da escala quando o número de pontos é reduzido. Neste caso, poder-se-á oferecer, extra escala, a opção de resposta *Não sei*.

Por outro lado, as escalas de resposta ímpares nem sempre são desejáveis. O objectivo do estudo pode definir como mais adequado o uso de uma "escolha forçada". É o caso de o estudo pretender prever a probabilidade de um voto ser no candidato A ou B. O inquérito deverá simular a situação de voto e impor uma resposta dicotómica, impondo ao inquirido a decisão que irá tomar junto das urnas.

Equilíbrio versus *uni-direccionalidade das escalas*

A direccionalidade possui um papel importante na elaboração da escala de resposta. Nos seus estudos, Peabody (1962, citado por Cummins & Gullone, 2000) concluem que somente 10% da componente de interpretação da escala pode ser atribuído à intensidade; o restante é atribuído à direcção. São os polos da escala contínua que determinam se uma escala é bidireccional (muito negativo a muito positivo) ou unidireccional (nada positivo a muito positivo) e a escolha está subjacente aos objectivos do estudo e ao conhecimento que o investigador tem da distribuição desta variável na população. A escolha de uma escala de resposta bidireccional para um questionário de satisfação, por exemplo, parte do princípio que existe pessoas insatisfeitas. Se tal não acontecer as características da distribuição das respostas será reduzida ao polo positivo da escala, com alguns "*outliers*" no seu polo negativo. Neste caso, a opção mais consistente é utilizar uma escala de resposta unidireccional e questionar apenas a intensidade da satisfação, ganhando sensibilidade na sua medida e garantindo uma distribuição mais consistente dos seus dados.

A mensuração de frequências e quantidades

A natureza da variável/dimensão a ser mensurada afecta o tipo de questão (aberta *vs.* fechada) e as opções escolhidas para as respostas fechadas, afectam a validade e a fiabilidade que conferem à medida.

Considere-se o caso em que a dimensão a aceder é uma "frequência". Ao perguntar a um indivíduo o número de vezes que exibiu um dado comportamento, estamos a colocar uma questão com formato

aberto. Ao perguntar, por exemplo, quantas vezes viu a versão original de um filme recente de uma forma espontânea as respostas podem ser "0, 1, 2, 3". Mas se perguntarmos quantas vezes viu "Música no Coração" surgem-nos problemas com esta pergunta "aberta". Isto porque quando a frequência do comportamento é elevada e dispersa no tempo, a fiabilidade da memória afecta a fiabilidade da medida. Entre as estratégias mais utilizadas para tentar ultrapassar o problema da grande dispersão das frequências, encontra-se o pedir ao inquirido que nos forneça apenas os valores associados a uma amostra reduzida do seu tempo de vida, como "no último dia", "na última semana", "no corrente ano". Ao utilizar-se uma estratégia como esta, não se deve ignorar que tal pode ter consequências na interpretação dos dados, já que a distribuição das frequências muda em função do intervalo a que estas se reportam (ver estudos de Winkielman, Knauper, & Schwarz, 1998).

Quando a questão colocada foca a dimensão "frequência", deve-se ter em conta a complexa tarefa cognitiva de *estimação de frequências.* É cognitivamente exigente e sujeita a vários tipos de enviesamentos. O investigador não pode ignorar este aspecto ao decidir usar este tipo de grandeza como "variável de estudo" (ver Conrad, Brown, & Cashman, 1998; Schwarz & Oyserman, 2001; Sudman, Bradburn, & Schwarz, 1996). Por exemplo, deve-se ter em conta que a amostra de eventos activos na nossa memória de trabalho afecta essa estimativa e que a forma como a questão está colocada pode afectar essa activação. É também relevante saber-se que os inquiridos tendem a sobrestimar a frequência de comportamentos raros e a subestimar a frequência de comportamentos frequentes (ver Sudman et al., 1996). As teorias pessoais sobre a distribuição da variável afectam as suas respostas. Por exemplo, ao reportar o consumo de álcool o inquirido que acredita que o álcool altera a estabilidade dos seus comportamentos, reporta o mesmo consumo independentemente do período de tempo referido enquanto o que acredita na instabilidade dos comportamentos, fornece uma resposta diferente consoante o período de tempo referido (ver Collins, Graham, Hansen, & Johnson, 1985). A estimação livre de frequência (resposta aberta) tem igualmente o problema de se estar a pedir a selecção de um número para representar a frequência. Estas respostas acabam por referir múltiplos de 5 e 10 (ver Tourangeau, Rips, & Rasinski, 2000) e valores prototípicos (por exemplo, quando a questão é "*Há quantos dias atrás*...", geralmente as respostas reportam valores associados à semana – 7 dias – ou ao mês – 30 dias; ver Huttenlocher, Hedges, & Bradburn, 1990). Daqui a vantagem de se usarem respostas de questão fechada, usando uma *escala de frequência*.

As *escalas de resposta de frequência* sendo truncadas, devem ser sustentadas em estudos piloto para que não se cometa o erro de oferecer aos inquiridos como alternativa de resposta apenas estimativas baixas ou estimativas elevadas. Para além de não permitir ao inquirido oferecer a sua resposta, o valor destas estimativas irá moldar a forma como o inquerido interpreta a questão colocada (Schwarz, Strack, Muller, & Chassein 1988; Winkielman et al., 1998). Por exemplo, se uma pergunta relativa a quão frequente é o inquirido sentir-se irritado (ex., com o seu chefe), for medida uma escala de baixas frequências (ex., de uma vez por ano a uma vez por mês), o inquirido irá interpretar a pergunta como se referindo a "grandes e significativas irritações". Mas se usarmos uma escala de elevada frequência (ex., uma vez por semana a várias vezes por dia), provavelmente o inquerido interpretará a questão como se referindo às pequenas desavenças que ocorrem no dia-a-dia. Múltiplos estudos demonstram esta influência das escalas de frequência nas respostas dos participantes relativamente a comportamentos de saúde (Wright, Gaskell, & O'Muircheartaigh, 1994), consumo de televisão (Schwarz, Hippler, Deutsch, & Strack, 1985) e comportamentos dos consumidores (Menon, Rhagubir, & Schwarz, 1995). Um exemplo claro é o que ocorre quando queremos saber a frequência de um sintoma de uma doença numa população psicossomática; quando a escala de resposta teve como âncoras 1 – menos de duas vezes por mês e 5 – várias vezes ao dia, 62% terem reportado valores acima de "duas vezes por mês" e quando tinha como âncoras 1 – nunca e 5 – Duas vezes por mês ou mais, apenas 39% reportaram valores acima de "duas vezes por mês" (Schwarz & Scheuring, 1992).

Em título de conclusão

Ao longo deste artigo foram apresentados diferentes tipos de resposta fechada a serem usadas num inquérito, e salientadas as consequências de o investigador decidir por um ou outro tipo. As escalas de avaliação contínua com 5 a 9 pontos parecem fornecer medidas intervalares, e fornecer medidas mais fiáveis, dependendo, porém, a sua validade da questão concreta colocada e dos rótulos (e mesmo números) utilizados em seus extremos (para o qual se recomendam pré-testes). Em geral as escalas de resposta gráficas permitem uma rápida compreensão da dimensão contínua da resposta e facilitam a resposta, isto sem comprometer as características métricas da medida que são semelhantes às das escalas de avaliação numérica.

Mas mais do que fornecer respostas ao investigador que pretende tomar uma decisão, este artigo fornece um conjunto de argumentos empíricos para sustentar as suas decisões. Mais concretamente, faz um resumo dos argumentos que se encontram dispersos na literatura do campo. Fica, assim, saliente o facto de que a escolha por um formato de resposta ou outro, não depende de questões de moda, estética, mas sim das implicações que esta escolha tem para uma válida e fiável operacionalização da medida. O investigador ao elaborar um inquérito deve ter em mente que este envolve uma interação com o inquirido que tem de compreender não só a questão colocada (ver Strack & Martin, 1987) mas a forma como lhe responder. Os trabalhos de Schwarz e Sudman (1996), têm chamado a atenção dos investigadores para a necessidade de perceberem a complexidade dos mecanismos cognitivos envolvidos nas respostas a um questionário. O investigador deve compreendê-los se quer perceber o significado das respectivas respostas. De entre estes processos cognitivos encontra-se, por exemplo, a compreensão da questão e a consequente "procura de significado". O investigador deverá ter em mente que o inquirido utiliza todos as características presentes num questionário para resolver qualquer ambiguidade na compreensão. Sem perceber como esta resolução de ambiguidade ocorre, o investigador pode cometer diversos erros de interpretação dos seus dados.

Alguns esclarecimentos adicionais

O presente artigo apenas foca o formato de respostas a serem inseridas num questionário. Deste modo, apesar de ser relevante para o campo de desenvolvimento de medidas complexas, o artigo não foca a natureza de itens a serem inseridos nessas medidas como escalas de atitudes contruídas com a metodologia proposta de Likert, escalas de aptidões etc.

Para um pensamento mais completo sobre a natureza do formato de itens a serem incluídos em medidas complexas, o leitor deve aceder a outro tipo de literatura. Uma literatura onde se destaca o contributo dos modelos desenvolvidos pela Teoria de Resposta ao Item (TRI), em complementaridade à análise de itens e ao estudo da fiabilidade e validade das escalas pela Teoria Clássica dos Testes (TCT). A TCT tem, por exemplo, referido que apesar de existirem influências do tipo de formato de uma resposta fechada na fiabilidade de uma escala compósita o mais relevante para a fiabilidade do instrumento é o número de itens utilizado nas escalas compósitas (Kline, 2005). Os modelos de TRI[3] usados sobre dados politómicos permitem descrever a probabilidade de um indivíduo responder a uma determinada categoria de um item, considerando as propriedades do item no todo e o nível do item no atributo (Hambleton, van der Linden, & Wells, 2010). Destas

[3] Dos diferentes modelos que foram desenvolvidos no âmbito da TRI para itens politómicos, destacam-se, o *Graded Response Model* e o *Multidimensional Model* de Samejima (1969), o *Rasch Rating Scale Model* de Andersen (1973) e Andrich, 1978, o *Partial Credit Model* de Masters (1982), e os *Non-Parametric Models* de Ramsey (1991) entre mais de 100 modelos de TRI identificados por van der Linden e Hambleton (1997).

abordagens surgem porém algumas recomendações sobre o formato da escala de resposta associada ao item, por exemplo recomendações para não usar, as escalas com categoria central, e as escalas bipolares. Adicionalmente os modelos de TRI sugerem a obtenção de escalas de intervalo a partir de respostas ordinais, eliminando a necessidade de desenvolver aparentemente categorias intervalares, quando este formato de resposta é incluído em medidas complexas. Assim, a TRI sugere uma análise específica do formato de resposta quando esta é associada a uma medida compósita. No entanto, apesar da TRI através dos seus diferentes modelos ajudar o investigador a conhecer o funcionamento de cada item, e ter uma palavra a dizer sobre o seu formato de resposta, as suas abordagens oferecem relevantes desafios técnicos para se conseguir o ajustamento dos modelos aos dados (e.g., o tamanho das amostras necessário à estabilidade do modelo; ver Hambleton et al., 2010).

Referências

Abend, R., Dan, O., Maoz, K., Raz, S., & Bar-Haim, Y. (2014). Reliability, validity and sensitivity of a computerized visual analog scale measuring state anxiety. *Journal of Behavior Therapy and Experimental Psychiatry*, *45*, 447-453.

Alwin, D. F. (1991). Research on survey quality. *Sociological Methods & Research*, *20*, 3-29.

Alwin, D. F. (1997). Feeling thermometers *versus* 7-point scales: Which are better?. *Sociological Methods & Research*, *25*, 318-340.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research*, *20*, 139-181.

Andersen, E. B. (1973). Conditional inferences for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31-44.

Andrews, F. M., & Withey, S. B. (1976). *Social indicators of well-being: The development and measurement of perceptual indicators.* New York: Plenum.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Averbuch, M., & Katzper, M. (2004). Assessment of visual analog *versus* categorical scale for measurement of osteoarthritis pain. *The Journal of Clinical Pharmacology*, *44*, 368-372.

Bartram, P., & Yelding, D. (1973). The development of an empirical method of selecting phrases used in verbal rating scales: A report on a recent experiment. *Journal of the Market Research Society*, *15*, 151-156.

Berman, D. R., & Stookey, J. A. (1980). Adolescents, television, and support for government. *Public Opinion Quarterly*, *44*, 330-340.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology*, *23*, 323-331.

Churchill Jr, G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, *21*, 360-375. Retrieved from https://doi.org/10.2307/3151463

Chyung, S. Y. 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.

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: A Monte-Carlo investigation. *Applied Psychological Measurement, 9,* 31-36.

Collins, L. M., Graham, J. W., Hansen, W. B., & Johnson, C. A. (1985). Agreement between retrospective accounts of substance use and earlier reported substance use. *Applied Psychological Measurement*, *9*, 301-309.

Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports*, *80*, 355-362.

Conrad, F. G., Brown, N. R., & Cashman, E. R. (1998). Strategies for estimating behavioural frequency in survey interviews. *Memory*, *6*, 339-366.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* Thousand Oaks, CA: Sage Publications, Inc.

Cork, R. C., Isaac, I., Elsharydah, A., Saleemi, S., Zavisca, F., & Alexander, L. (2004). A comparison of the verbal rating scale and the visual analog scale for pain assessment. *The Internet Journal of Anesthesiology*, *8*, 23-38.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, *24*, 227-245.

Cox III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*, 407-422.

Cummins, R. A., & Gullone, E. (2000, March). Why we should not use 5-point Likert scales: The case for subjective quality of life measurement. *Proceedings of 2nd International Conference on Quality of Life in Cities* (pp. 74-93). Singapore: National University of Singapore.

Flynn, D., van Schaik, P., & van Wersch, A. (2004). A comparison of multi-item Likert and visual analogue scales for the assessment of transactionally defined coping function. *European Journal of Psychological Assessment*, *20*, 49-59.

Funke, F. (2016). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, *34*, 244-254.

Funke, F., & Reips, U. D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, *24*, 310-327.

Givon, M. M., & Shapira, Z. (1984). Response to rating scales: a theoretical model and its application to the number of categories problem. *Journal of Marketing Research*, *21*, 410-419.

Grapentine, T. (2003). Scales: Still problematic 10 years later. *Marketing Research, 15*, 45-46.

Hambleton, R. K., van der Linden, W. J., & Wells, C. S. (2010). IRT Models for the Analysis pf Polytomously Scored Data: Brief and Selected History of Model Building Advances. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 21-42). New York, NY: Routledge, Taylor & Francis Group.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society*, *159*, 445-492. Retrieved from https://doi.org/10.2307/2983326

Hartley, J., Trueman, M., & Rodgers, A. (1984). The effects of verbal and numerical quantifiers on questionnaire responses. *Applied Ergonomics*, *15*, 149-155.

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, *18*, 98-99.

Hofmans, J., & Theuns, P. (2008). On the linearity of predefined and self-anchoring Visual Analogue Scales. *British Journal of Mathematical and Statistical Psychology*, *61*, 401-413.

Huskisson, E. C. (1974). Measurement of pain. *The Lancet*, *304*(7889), 1127-1131.

Huttenlocher, J., Hedges, L. V., & Bradburn, N. M. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 196-213.

Jenkins, G. D., & Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology*, *62*, 392-398.

Kennedy, R., Riquier, C., & Sharp, B. (1996). Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting Measurement and Analysis for Marketing*, *5*, 56-70.

Kiess, H. O., & Bloomquist, D. W. (1985). *Psychological research methods: A conceptual approach.* Needham Heights, MA: Allyn & Bacon.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: SAGE Publications, Inc.

Krosnick, J. A., & Alwin, D. F. (1989). Aging and susceptibility to attitude change. *Journal of Personality and Social Psychology*, *57*, 416-425.

Krosnick, J. A., & Presser, S. (2010). Questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed.). West Yorkshire, England: Emerald Group.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 5-55.

Lindzey, G. E., & Guest, L. (1951). To repeat-check lists can be dangerous. *Public Opinion Quarterly, 15*, 355-358.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

McCormack, H. M., David, J. D. L., & Sheather, S. (1988). Clinical applications of visual analogue scales: A critical review. *Psychological Medicine*, *18*, 1007-1019.

McKelvie, S. J. (1978). Graphic rating scales – How many categories?. *British Journal of Psychology*, *69*, 185-202.

Meek, P. M., Sennott-Miller, L., & Ferketich, S. L. (1992). Focus on psychometrics scaling stimuli with magnitude estimation. *Research in Nursing & Health*, *15*, 77-81.

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral frequency judgments: An accessibility-diagnosticity framework. *Journal of Consumer Research*, *22*, 212-228.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*, 398-407. Retrieved from https://doi.org/10.1037/0033-2909.100.3.398

Moxey, L. M., & Sanford, A. J. (1992). Context effects and the communicative functions of quantifiers: Implications for their use in attitude research. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 279-296). New York, NY: Springer.

Myers, J. H., & Warner, W. G. (1968). Semantic properties of selected evaluation adjectives. *Journal of Marketing Research*, *5*, 409-412.

Myles, P. S., & Urquhart, N. (2005). The linearity of the visual analogue scale in patients with severe acute pain. *Anaesthesia and Intensive Care*, *33*, 54-58.

Myles, P. S., Troedel, S., Boquest, M., & Reeves, M. (1999). The pain visual analog scale: Is it linear or nonlinear?. *Anesthesia & Analgesia*, *89*, 1517-1520. Retrieved from https://doi.org/10.1213/00000539-199912000-00038

Neuman, L., & Neuman, Y. (1981). Comparison of six lengths of rating scales: Students attitude toward instruction. *Psychological Reports, 48*, 399-404.

Noelle-Neumann, E. (1970). Wanted: Rules for wording structured questionnaires. *Public Opinion Quarterly*, *34*, 191-201. Retrieved from https://doi.org/10.1086/267789

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*, 625-632.

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills*, *68*, 549-550.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. M. (1957). *The measurement of meaning.* Urbana: University of Illinois Press.

Osinski, I. C., & Bruno, A. S. (1998). Categorías de respuesta en escalas tipo Likert. *Psicothema*, *10*, 623-631.

Paul-Dauphin, A., Guillemin, F., Virion, J. M., & Briançon, S. (1999). Bias and precision in visual analogue scales: A randomized controlled trial. *American Journal of Epidemiology*, *150*, 1117-1127.

Peabody, D. (1962). Two components in bipolar scales: Direction and extremeness. *Psychological Review*, *69*, 65-73. Retrieved from https://doi.org/10.1037/h0039737

Pepper, S. (1981). Problems in the quantification of frequency expressions. In D. W. Fiske (Ed.), *Problems with language imprecision: New directions for methodology of social and behavioral science* (Vol. 9, pp. 25-41). San Francisco: Jossey-Bass.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1-15.

Ramsay, J. (1991). Kernal smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611-630.

Rasmussen, J. L. (1989). Analysis of Likert-scale data: A reinterpretation of Gregoire and Driver. *Psychological Bulletin, 105*, 167-170.

Rausch, M., & Zehetleitner, M. (2014). A comparison between a visual analogue scale and a four-point scale as measures of conscious experience of motion. *Consciousness and Cognition*, *28*, 126-140. Retrieved from https://doi.org/10.1016/j.concog.2014.06.012

Reips, U. D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in internet-based research: VAS Generator. *Behavior Research Methods*, *40*, 699-704.

Revill, S. I., Robinson, J. O., Rosen, M., & Hogg, M. I. J. (1976). The reliability of a linear analogue for evaluating pain. *Anaesthesia*, *31*, 1191-1198.

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

Schriesheim, C. A., & Castro, S. L. (1996). Referent effects in the magnitude estimation scaling of frequency expressions for response anchor sets: An empirical investigation. *Educational and Psychological Measurement*, *56*, 557-569.

Schriesheim, C. A., & Novelli Jr, L. (1989). A comparative test of the interval-scale properties of magnitude estimation and case III scaling and recommendations for equal-interval frequency response anchors. *Educational and Psychological Measurement*, *49*, 59-74.

Schriesheim, C., & Schriesheim, J. (1974). Development and empirical verification of new response categories to increase the validity of multiple response alternative questionnaires. *Educational and Psychological Measurement*, *34*, 877-884.

Schuman, H., & Scott, J. (1987). Problems in the use of survey questions to measure public opinion. *Science*, *236*(4804), 957-959.

Schwarz, N., & Hippler, H.-J. (1990). *Response alternatives: The impact of their choice and presentation order*. (ZUMA-Arbeitsbericht, 1990/08). Mannheim: Zentrum für Umfragen, Methoden und Analysen – ZUMA. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-67257

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*, 388-395.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570-582.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *The American Journal of Evaluation*, *22*, 127-160.

Schwarz, N., & Scheuring, B. (1992). Frequency-reports of psychosomatic symptoms: What respondents learn from response alternatives. *Zeitschrift fur Klinische Psychologie, 21*, 197-208.

Schwarz, N., Strack, F., Müller, G., & Chassein, B. (1988). The range of response alternatives may determine the meaning of the question: Further evidence on informative functions of response alternatives. *Social Cognition*, *6*, 107-117.

Schwarz, N. E., & Sudman, S. E. (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Francisco, CA: Jossey-Bass.

Srinivasan, V., & Basu, A. K. (1989). The metric quality of ordered categorical data. *Marketing Science*, *8*, 205-230.

Stevens, S. S. (1975). *Psychophysics*. New Jersey: Transaction Publishers.

Strack, F., & Martin, L. L. (1987). Thinking, judging, and communicating: a process account of context effects in attitude surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology: Recent research in psychology*. New York, NY: Springer. Retrieved from https://doi.org/10.1007/978-1-4612-4798-2_7

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco, CA: Jossey-Bass.

Svensson, E. (2000). Comparison of the quality of assessments using continuous and discrete ordinal rating scales. *Biometrical Journal*, *42*, 417-434.

Swait, J., & Adamowicz, W. (2001). The influence of task complexity on consumer choice: A latent class model of decision strategy switching. *Journal of Consumer Research*, *28*, 135-148.

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

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. B. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology. Building a bridge between disciplines: Report of the advanced research seminar on cognitive aspects of survey methodology* (pp. 73-100). Washington, DC: National Academy Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96*, 394-401. Retrieved from https://doi.org/10.1037/0033-2909.96.2.394

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Viswanathan, M., Sudman, S., & Johnson, M. (2004). Maximum *versus* meaningful discrimination in scale response: Implications for validity of measurement of consumer perceptions about products. *Journal of Business Research*, *57*, 108-124. Retrieved from https://doi.org/10.1016/S0148-2963(01)00296-X

Ware, J. E., & Gander, B. (1994). The SF-36 health survey: Development and use in mental health research and the IQOLA project. *International Journal of Mental Health, 23*, 49-73. Retrieved from https://doi.org/10.1080/00207411.1994.11449283

Weathers, D., Sharma, S., & Niedrich, R. W. (2005). The impact of the number of scale points, dispositional factors, and the status quo decision heuristic on scale reliability and response accuracy. *Journal of Business Research*, *58*, 1516-1524.

Wewers, M. E., & Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health*, *13*, 227-236.

Wildt, A. R., & Mazis, M. B. (1978). Determinants of scale response: Label *versus* position. *Journal of Marketing Research*, *15*, 261-267.

Wills, C. E., & Moore, C. F. (1994). Focus on psychometrics. A controversy in scaling of subjective states: Magnitude estimation *versus* category rating methods. *Research in Nursing & Health*, *17*, 231-237.

Winkielman, P., Knäuper, B., & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of emotion frequency questions. *Journal of Personality and Social Psychology*, *75*, 719-728.

Worcester, R. M., & Burns, T. R. (1975). Statistical examination of relative precision of verbal scales. *Journal of the Market Research Society*, *17*, 181-197.

Wright, D. B., Gaskell, G. D., & O'Muircheartaigh, C. A. (1994). How much is 'quite a bit'? Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology*, *8*, 479-496.

The closed-ended questions formats: The nature, validity and reliability of their responses

This paper aims to support researchers in the decisions about the use of closed answer questions in their questionnaires by critically reviewing the literature regarding the implications of such decisions for the nature, validity and reliability of the measure. This literature's review provides arguments for the researcher to decide on how best to operationalize their variables through a closed response. Arguments presented support decision-making in the construction of response options to be provided to the respondent, and the type of scale to be used: graphical or non-graphical; categories or continuous assessment; with 3 or more points; with or without labels and in this case, with what kind of labels, etc. Other topics are illustrated to be considered for construction of a response scale analyzing the specific case in which "perceived frequencies" are measured.

Key words: Assessment scales, Questionnaires, Bias.

Submissão: 15/11/2018 *Aceitação:* 02/07/2019