Disponível em
http://www.anpad.org.br/rac

RAC, Rio de Janeiro, v. 19, 2ª Edição Especial, art. 3,
pp. 157-177, Agosto 2015
http://dx.doi.org/10.1590/1982-7849rac20151559

Mensuração de Atitude: Proposição de um Protocolo de Elaboração de Escalas

Measurement of Attitude: Proposition of a Protocol for Preparation of Scales

Rafael Lucian
Faculdade Boa Viagem – MPGE/FBV/DeVry

Jairo Simião Dornelas
Universidade Federal de Pernambuco – UFPE/CCSA/DCA

Artigo recebido em 23.07.2013. Última versão recebida em 19.05.2014. Aprovado em 21.05.2014. Publicado online em 19.11.2014.

Resumo

Este ensaio teórico dedicou-se a estudar como escalas são formadas e a partir de que procedimentos é possível considerá-las válidas e aptas para o uso como instrumento científico legítimo. Nesta ótica, o objetivo deste artigo foi propor um protocolo de construção de escalas de mensuração de atitude. O protocolo proposto configura-se como a reunião lógica de passos baseados em teóricos como Allport e Hartman (1925), Thurstone (1928), Likert (1932), Campbell e Fiske (1959) e Bock (1972), que permeiam todo o caminho da elaboração de escalas, quais sejam a definição de Construto, a escolha da escala em si, a elaboração dos itens, a purificação da escala e, finalmente, a validação desta. Ao final do estudo, apresenta-se um protocolo de elaboração de escalas específico para mensuração de atitude que se diferencia dos protocolos vigentes de Churchill (1979), Rossiter (2002) e DeVellis (2003) ao reunir ineditamente um conjunto de técnicas promissoras como, principalmente, a delimitação objetiva do constructo por grupo focal, proposição de uma escala em si dicotômica, purificação da escala por teoria de resposta ao item (TRI) e validação preditiva.

Palavras-chave: mensuração de atitude; protocolo de elaboração de escala; mensuração escalar; escala múltipla.

Abstract

This theoretical essay aims to study how scales are developed and through which procedures they can be considered valid and suitable for use as legitimate scientific instruments. In this perspective, this paper's objective was to develop a protocol for constructing scales to measure attitude. The proposed protocol is configured as a logical meeting of steps based on theorists such as Allport and Hartman (1925), Thurstone (1928), Likert (1932), Campbell and Fiske (1959) and Bock (1972), which permeate all aspects of drafting scales, including construct definition, the choice of the scale itself, item preparation, scale purification, and finally its validation. At the end of the study, we present a protocol for the preparation of specific scales to measure attitude that differs from existing protocols in Churchill (1979), Rossiter (2002) and DeVellis (2003). This is in order to unite for the first time a set of promising techniques, primarily the objective delineation of the construct using focus group methodology, the proposition of an inherently dichotomous scale, scale purification via item response theory (IRT), and predictive validity.

Key words: attitude measurement; scale-proposition protocol; scale measurement; multiple scale.

Introdução

A sociedade humana pode ser investigada sob diversas perspectivas. Nas ciências sociais, há forte interesse em escrutinar assuntos relevantes às pessoas, incluindo as formas como se organizam, tomam decisões, lidam com tecnologia, constroem conhecimento e têm seu comportamento mensurado de modo empírico.

De fato, no decurso do tempo, o comportamento se caracterizou por interferir nas decisões e grande parte do interesse científico na área social concentrou-se em estudar esse comportamento. Na administração, não foi diferente, pois comportamento é a base para o estudo das organizações. Constata-se, então, que as pessoas têm comportamento sócio-organizacional-grupal que, cada vez mais, deseja-se ser conhecido. Neste segmento particular, uma área consagrada, porém ainda promissora para estudos, refere-se a compreender como as pessoas reagem diante de situações conhecidas, ou seja, suas atitudes.

Pessoas se organizam em grupos ou redes com diferentes fins, assim, entender os motivos e as ações dessas organizações é de particular interesse dos estudiosos. Os pesquisadores sociais buscam explicações e entendimento sobre os diversos aspectos da vida das pessoas em organizações e, para tal, fazem uso deliberado de metodologias científicas e de mensurações empíricas.

As pessoas, grupos e organizações, quando necessitam mensurar atitude, buscam formas de fazê-lo com o menor desperdício de recursos e maior precisão possível. Para tanto, ao invés de delinear pesquisas para cada evento de mensuração desejado, buscam modelos prontos e acreditados pela academia que transpareçam confiabilidade, como, por exemplo, em Zemack-Rugar, Corus e Brinberg (2012), Lee, Cornwell e Babiak (2012), Pérez e Bosque (2013), Know *et al*. (2013) e Bauerband e Galupo (2014). O impacto, então, do protocolo de elaboração é fornecer um instrumentário que permita economizar recursos no planejamento da pesquisa e que seja capaz de gerar escalas válidas e confiáveis.

Nesta perspectiva, vários modelos, teorias e Construtos foram criados com o propósito de prover um suporte razoável para as tomadas de decisão (Pooja & Sagar, 2012), incluindo a mensuração de atitude, um campo quase centenário de estudos que busca identificar e prever, através de um instrumento escalar, qual é o conjunto de comportamentos predefinidos de uma pessoa sobre algo (Sanches, Meireles, & Sordi, 2011).

Atitude é uma predisposição, relativamente estável e organizada, para reagir na forma de opiniões ou de atos em presença de objetos, de maneira determinada (Bardin, 2009), que representa uma posição mental consistente, manifesta, sobre algo ou alguém (Ander-Egg, 1978).

Para uma melhor compreensão do fenômeno investigado, é importante diferenciar atitude, intenção e comportamento. Atitude é a expressão do sentimento em relação a algo; enquanto intenção é a propensão declarada a fazê-la; e comportamento consiste na ação em si. Logo, nesta perspectiva, compreender a atitude é importante, pois em algum grau ela explica e prevê o comportamento das pessoas (Bagozzi, 1981).

Assim, por ser um constructo psicológico, só é possível acessar a atitude de uma pessoa se ela comunicá-la ou demonstrá-la, pois atitude é essencialmente uma disposição mental em face de uma ação potencial (Mann, 1970). Nesta perspectiva, embora o comportamento e a atitude sejam Construtos diferentes, eles estão relacionados (Bagozzi, 1981) e, por isso, é possível mensurar a atitude através da observação do comportamento das pessoas em relação a algo conhecido e determinado, como no clássico experimento de Grim (1936).

Se a observação dos atos é um procedimento que demanda mais tempo e praticamente inviabiliza estudos em grande escala, a mensuração das opiniões (expressão oral ou escrita da atitude) oferece diversas vantagens em relação à economia de recursos. Talvez, por isso que vem sendo conferida, desde Galton (1880), grande importância aos estudos de mensuração por meio de afirmações escritas.

Mensuração, por sua vez, segundo Crowther (1995), é uma técnica que faz uso de instrumentos de precisão para se medir qualidades desejadas com base numérica. Portanto, a princípio, qualquer coisa observável direta ou indiretamente, incluindo a atitude, pode ser mensurável desde que se tenha um instrumento apropriado para tal.

Contudo o processo de mensuração é mais amplo do que a atribuição de números aos objetos que representem quantitativamente algum atributo que se queira mensurar; seu objetivo é prover um mecanismo de análise que gere informação e sirva de fomento para uma tomada de decisão inteligente (Pooja & Sagar, 2012).

Os pioneiros nesta linha, Allport e Hartman (1925), sugeriram a mensuração de atitude sob duas dimensões: o sentido e a intensidade. Para eles, era possível investigar qual era a atitude de uma pessoa (positiva ou negativa) e, além disso, qual a sua intensidade.

Esse movimento inicial de proposição de escalas como método de mensuração em ciências sociais e psicologia foi inspirado na física. Allport e Hartman (1925) e Thurstone (1928) se basearam na lógica das escalas métricas para proporem um instrumento de mensuração bastante similar. Essa instrumentação derivou do desejo de se construir uma ferramenta capaz de comparar grupos, ao contrário das possibilidades de medições anteriores.

O principal avanço nos estudos de mensuração de atitude, todavia, foi a proposição original de Likert (1932), que sugeriu uma escala unificada em que através do mesmo instrumento fosse possível identificar o sentido e a intensidade da atitude. Desde então (até os dias atuais), a mensuração neste formato é a mais aceita entre os pesquisadores e profissionais de mercado (Sanches *et al.*, 2011).

Um aspecto da mensuração que ganhou grande impulso a partir das ideias de Likert (1932) foi a validação de escala. Esta corrente de estudos surgiu para responder à principal questão da área, que era como saber se a escala elaborada tinha a capacidade de mensurar o Construto desejado. Nesta ótica, validação com uso de técnicas estatísticas é o nome dado ao conjunto de procedimentos utilizados para conferir maior credibilidade ao processo de mensuração. Entre as principais contribuições neste segmento, destaca-se o trabalho de Campbell e Fiske (1959).

Com os avanços da etapa de validação, o foco da academia voltou-se para o desenvolvimento de protocolos de elaboração de escalas, descritivos e explicativos, que necessitam substancialmente da robustez de método e de técnicas de pesquisa. Neste sentido, a elaboração de escalas de mensuração envolve a construção de um instrumento em si e a associação de conceitos qualitativos com as métricas quantitativas, ou seja, a atribuição de números a objetos segundo alguma regra determinada (Pooja & Sagar, 2012), a qual busca disciplinar o estudo do fenômeno.

Com tal direcionamento, um protocolo de elaboração de escalas é um conjunto organizado de etapas a cumprir, com o uso adequado de técnicas selecionadas, para se construir uma escala de mensuração válida (Rossiter, 2002). A tarefa de construção de protocolo é uma atividade que permeia as diversas áreas da ciência (Churchill, 1979), mesmo que de forma pouco recorrente e sem utilização de métodos específicos.

Não obstante todo esforço para elaboração de protocolos, Straub (1989) manifestou preocupação, pois as pesquisas, além de não adotarem um protocolo como referência comum, também, não aplicavam, em sua maioria, qualquer protocolo para criação de escalas. Ademais, foi constatado, no mesmo estudo, que 83% dos trabalhos incluídos em sua amostra, e que usaram escala, não adotaram qualquer critério de validação destas, comprometendo, portanto, todas as análises efetuadas por se basearem em informações provenientes de instrumentos não validados.

Esses números são ainda piores no estudo de Kaptein, Nass e Markopoulos (2010), em que foi identificado um total de 92% de trabalhos investigados que realizavam mensurações fora dos padrões estipulados pelos protocolos. O descrédito sobre a validade e confiabilidade das escalas também está representado pelo estudo de Turner e Zolin (2012), que, ao depararem-se com um cenário de elaboração

livre de escalas, propuseram-se a investigar quais dos instrumentos publicados pela literatura eram realmente aptos a mensurar o constructo por eles investigado.

Mais especificamente, Doll e Torkzadeh (1991), também, indicaram equívocos na mensuração da satisfação do usuário final, tal qual Bagozzi (1981), que igualmente levantou a suspeita de que a estatística utilizada em seu tempo não era apropriada e os resultados imprecisos. Nesta mesma perspectiva, Petter, Rai e Straub (2012) enfatizam que, em situações como estas, há necessidade de se propor um conjunto de regras que norteie os estudos que desejam fazer medições por escalas.

Assim, ao perceber esse vácuo, que só tem crescido nos últimos tempos, no que concerne ao uso não totalmente consistente de mensurações de atitudes por escalas construídas ou adaptadas, sem os ritos próprios de elaboração e validação, desta forma, produzindo resultados num contexto sem a necessária fidedignidade metodológica, é que surge a ideia de formular, com o uso de diversas técnicas, um protocolo de desenvolvimento de escalas.

Nesta perspectiva que nasce a oportunidade de pesquisa adotada por este estudo, qual seja propor um protocolo de construção de escalas de mensuração de atitude. Entretanto, para ter êxito em tal tarefa, é necessário revisar e discutir a teoria dos principais pontos relativos ao processo de elaboração de protocolos de mensuração e aprofundar tais conceitos no campo do estudo das atitudes.

Concretiza-se tal fim pelo caminho percorrido pelos próximos tópicos que abordam passo a passo as etapas de elaboração de escalas de mensuração de atitude e culminam com a apresentação do protocolo proposto e sua comparação com os modelos existentes.

Descrição do Protocolo

O protocolo proposto foi construído com o objetivo de apresentar os passos lógicos para a construção de escalas de mensuração de atitude. É de ordem incremental e fruto do conhecimento e análise crítica dos principais estudos neste campo.

As próximas seções se dedicam a apresentar a discussão suscitada sobre cada uma das etapas de um protocolo de elaboração de escalas. A organização do texto se deu de tal forma que, para cada ponto, é apresentada sua discussão teórica e, ao final, apresenta-se o esquema do protocolo proposto.

Definição do construto

As pessoas são sensíveis aos estímulos que recebem das coisas e chama-se empírica toda intuição que se relaciona ao objeto por meio da experiência. Quando este objeto é indeterminado, denomina-se fenômeno (Kant, 2009).

Em um fenômeno, chama-se de matéria (objeto) aquilo a que as sensações se dirigem. Atributo, por sua vez, é definido, por Rossiter (2002), como as características independentes e observáveis do objeto. Neste raciocínio, os atributos são partes de um objeto que, por sua vez, compõem o fenômeno, sendo todos esses limitados à possibilidade da experiência. Assim, é importante entender que o processo de identificação do Construto requer obrigatoriamente as definições de fenômeno, objeto e atributo (Chapa & Stringer, 2013).

Construto, como definido por Edwards e Bagozzi (2000), é um termo conceitual utilizado para descrever teoricamente um fenômeno de interesse. Rossiter (2002) afirma que um Construto deve ser conceitualmente definido em termos de objeto, atributo e população. As questões básicas são: o que é o objeto e de que ele é composto; quais são seus atributos e de que eles são compostos; e por quem é formada a população que irá responder à enquete.

Outrora, Guttman (1943), também, definira Construto como conjunto de atributos, ou uma ampla classe de comportamentos. Nesta ótica, na interpretação de Lee *et al*. (2012), para que um Construto verdadeiro se forme, todos os atributos ou classes de comportamentos devem estar unidos por algum critério comum que justifique a sua classificação em um mesmo conjunto. Assim, percebeu-se que a definição do Construto precisava ser muito precisa, deixando claro o que está incluso e o que está excluso dela (Churchill, 1979; DeVellis, 2003), sendo esta uma das grandes dificuldades das pesquisas em ciências sociais em que muitos Construtos são abstrações teóricas, inobserváveis.

Bearden e Netemeyer (1999) ressaltam ainda que é necessário que a escala de atitude para mensurar Construtos sociais esteja de acordo com alguma teoria e que seus itens sejam correspondentes aos Construtos teorizados. Em síntese, ele é formado, então, por um objeto (devidamente delimitado), seus atributos (conjunto de comportamentos escolhidos) e pelos respondentes (universo).

No entanto os procedimentos de identificação de Construto carregam certa subjetividade, posto que é uma construção de ordem teórica. Chapa e Stringer (2013) argumentam que o uso de técnicas que envolvam subjetividade é permitido desde que haja um mínimo rigor para se manter conforme a proposta epistemológica da mensuração por escalas.

Há, contudo, um esforço no sentido de objetivar essa tarefa por parte de Rossiter (2002), que oportunamente apresentou alguns questionamentos guias para esta identificação. Seguindo esses questionamentos, a primeira fase do protocolo consiste em responder às perguntas, as quais também servem para nortear a elaboração dos itens.

Os questionamentos de Rossiter (2002) devem ser utilizados para cumprir a etapa de definição do Construto. Quais sejam: Qual o Construto que será estudado? Quais são os limites deste Construto? Quais são as manifestações observáveis do Construto? Quais são os objetos observáveis do Construto? Quais são os atributos de cada objeto observável do Construto? Qual público se pretende ter como respondente?

Através dos questionamentos supracitados é possível delimitar o que será estudado e qual será o público alvo. Para tal, devem ser realizadas entrevistas pessoais estruturadas com o público-alvo.

Definir a escala em si

Inicialmente, para um melhor entendimento, é necessário definir o termo escala em si. Pragmaticamente, é comum utilizar o termo escala tanto para definir o instrumento de mensuração quanto seu formato. Por exemplo, quando Malhotra (2011) utiliza o termo escala Likert, ele se refere ao formato, enquanto Parasuraman, Zeithaml e Berry (1985), ao apresentarem sua escala SERVQUAL, fazem-no ao instrumento de mensuração. Por conseguinte, para evitar tal duplicidade semântica, adota-se o termo escala em si exclusivamente para referenciar-se ao formato.

Até onde se pôde apurar na literatura especializada em protocolos de desenvolvimento de escalas, não foi encontrado tratamento específico para a etapa de escala em si, sendo esmagadora a presença da escala em si nos moldes da escala de Likert, como, por exemplo, em Churchill (1979), Rossiter (2002) e Devellis (2003).

Neste artigo, propõe-se que essa escolha não seja obrigatória e apresenta-se uma inovação em relação à escala em si. Com base em Likert (1932), iniciou-se uma revisão teórica com o objetivo de identificar seus pontos de melhoria e, ao final, propor uma escala em si revisada que minimize tais fragilidades.

Conquanto a escala em si de Likert (1932) seja utilizada em diversas áreas, ela foi elaborada originalmente para o Construto atitude. A escala fora teorizada considerando que a atitude não poderia ser captada por um único item (propondo então a escala multi-itens) e teve desenvolvida uma forma de se mensurar simultaneamente o sentido e a intensidade desta atitude.

Com a popularização da escala de Likert, os debates sobre seus aspectos se intensificaram e um dos pontos mais explorados foi a importância do ponto neutro. Komorita (1963), um dos principais teóricos quanto a esse aspecto, sugere que não é possível definir claramente um ponto neutro em escalas ordinais, que é como classifica a escala em discussão. Anteriormente, Peabody (1962) e Sjoberg e Nett (1968) também já afirmaram que a presença ou ausência de uma categoria neutra é indiferente para a validação da escala. Portanto, a decisão por manter ou retirar o neutro deve ser tomada de acordo com a necessidade do pesquisador.

Outro ponto de discussão sobre as escalas em si é relativo aos seus rótulos. Desde o início, têm sido utilizados palavras e números para tal, embora boa parte dos críticos questione essa escolha, como Boyd, Westfall e Stasch (1977).

Diante deste impasse teórico, Derham (2011) realizou uma série de testes empíricos, então, observando o comportamento de três tipos de escalas em si do tipo Likert, sendo a primeira utilizando apenas palavras nos rótulos, a segunda utilizando série numérica para indicar os graus da escala em si e a terceira completamente gráfica. Os resultados do estudo indicaram que o formato mais confortável ao respondente seria palavras como rótulos. Este formato apresentou melhor desempenho em seis dos sete atributos testados.

À parte ao formato dos rótulos, o número de graus na escala em si tem despertado grande interesse da academia e até hoje não há consenso sobre seu efeito na mensuração de atitude. As escalas em si do tipo Likert carregam dois componentes: direção e intensidade. Contudo, como já antevia Cronbach (1951), restariam dúvidas sobre a efetividade da mensuração sobre intensidade.

Sob tal perspectiva, uma preocupação referente à mensuração de intensidade e ao número de itens da escala é relativa ao não balanceamento dos modelos politômicos (Nunnally, 1978). E quando de uma escala de cinco pontos, os intervalos negativos tendem a ser maiores que os positivos, e este comportamento de desbalanceamento independe do Construto que se esteja mensurando corretamente (Tomas & Oliver, 1999).

Este comportamento assimétrico entre positivo e negativo é explicado, ainda, por Rozin e Royzman (2001), que apontam para o fato de as avaliações negativas serem mais fortes, intensas e rápidas que as positivas. Assim, o conjunto de várias percepções positivas contra apenas uma negativa pode resultar em atitude negativa, não respeitando a lógica aritmética que suporta o modelo Likert. Uma alternativa para tal é a dicotomização da escala, que se justifica pela melhor assimetria entre positividade e negatividade (Anderson, 1965).

Sobre essa modificação no número de graus da escala em si, Komorita (1963) concluiu, por meio de investigações empíricas, que a confiabilidade da escala independe do número de alternativas de resposta. Na última obra citada, constatou-se que escalas em si dicotômicas e politômicas tendem a ter o mesmo grau de confiabilidade quando comparadas.

Rodriguez (2005), por meio de metanálise, concluiu que uma escala em si com três opções de resposta é suficiente, sendo uma positiva, uma negativa e uma neutra. Este autor destaca que o efeito da diminuição do número de graus de escolha encolhe o teste e, proporcionalmente, aumenta sua eficiência para grandes quantidades de respondentes. Em complemento, denota que o tempo gasto na resposta do questionário é proporcional ao número total de alternativas e o uso de escala em si tricotômica diminui o tempo na coleta de informação.

Viswanathan, Sudman e Johnson (2004) demonstram preocupação com a relação entre a escala em si e os testes estatísticos, visto que a definição do número de itens afetará os testes estatísticos a serem realizados. De fato, o uso de uma escala com muitos pontos pode não prover uma base de dados válida para a realização de inferências estatísticas, visto que, de acordo com o tamanho da amostra, pode resultar em uma dispersão entre os respondentes, limitando o uso de alguns testes estatísticos (Lake, 2014). Além do que, testes mais modernos como a teoria de resposta ao item se comportam melhor com escalas em si reduzidas (Bock, 1972).

Assim, a proposição de escala em si deste artigo é de utilizar um formato dicotômico, além de um ponto neutro com poder de anular a questão em caso de não aplicação ou indecisão do respondente e rótulos como palavras. Ao tornar a escala em si de caráter nominal, o mecanismo torna-se mais preciso em relação à mensuração do sentido da atitude do que o modelo original de sentido e intensidade.

A questão da intensidade foi descartada da escala em si proposta também pelo fato desta característica ter sido uma proposição de Allport e Hartman (1925) com fins de mensuração, enquanto que a teoria de atitude não faz uso desta perspectiva, como, por exemplo, Bagozzi (1981), que afirma as dimensões da atitude positiva ou negativa em seu estudo clássico de hipóteses, porém não enfatiza a intensidade.

Uma vez definida a escala em si, é possível iniciar a elaboração dos itens da escala, os quais devem ser projetados especificamente para esta escala em si, como será discutido na próxima seção.

Elaboração dos itens da escala

A elaboração dos itens é o terceiro passo do protocolo proposto. Nesta fase, foi adotada a proposição de Allport e Hartman (1925), por ser a mais completa e ter sido base de todas as outras conhecidas.

Por essa estratégia de elaboração, inicialmente, é necessário obter opiniões do público-alvo, pois elas são a base para redação dos itens que irão compor a escala. Para tanto, esta coleta deve ocorrer através de um levantamento com uso de questionário (utilizando perguntas do tipo: qual a sua opinião sobre tal coisa), e, após esta fase ser superada, deve-se fazer uso de um grupo focal com especialistas para que eles promovam a seleção, entre todas as opiniões coletadas, das que serão úteis para compor os itens da escala. É enfatizado por Likert (1932) que coletar opiniões diretamente com o público-alvo elimina a quase insuperável barreira do pesquisador tentar compor afirmações com o vocabulário e estilo textual dos respondentes.

Os itens, então, são inicialmente elaborados a partir da coleta de opiniões por levantamento de uma amostra da população desejada. Deve-se coletar o maior número de opiniões possível, já que nenhum estudo anterior fala na quantidade exata. Isto deve ser feito até o ponto em que o pesquisador note que a coleta não está mais contribuindo para o acréscimo de itens (ou seja, saturou-se), quando se deve decidir por suspender a coleta.

Finalmente, após ter os dados do levantamento tabulados, o pesquisador deve organizar um grupo focal com especialistas para que eles selecionem, entre todas as opiniões expostas, as que devem ser incluídas no questionário. Os critérios que os especialistas precisam levar em conta para esta seleção são a relevância e aderência da opinião (que se tornará uma afirmação da escala) com o constructo estudado. Dada a complexidade desta intervenção, não deve-se fixar numericamente a quantidade de afirmações finais desejadas, porém, quanto mais específico for o constructo investigado, menor será o conjunto natural de frases finais. Como nenhuma afirmativa poderá ser incluída na escala nos passos posteriores, o pesquisador deve precaver-se para não partir para purificação com um conjunto demasiadamente enxuto sob pena de não ter seu constructo devidamente representado na escala final.

O passo seguinte à elaboração da escala é realizar uma validação de face. Nesta etapa, observaram-se as sugestões de Hardesty e Bearden (2004) e os itens devem ser apresentados aos público-alvo para julgar se a escala proposta parece eficaz para mensurar o Construto intencionado, como é discutido na próxima subseção.

Purificação da escala

Após a definição de uma versão preliminar da escala, faz-se necessário promover a validação de face. No protocolo desta pesquisa, esta fase tem o objetivo de observar a concordância do grupo de especialistas convidado com a capacidade da escala de mensurar o Construto pretendido, em primeira etapa, bem como os subsídios para se efetuar o cálculo da confiabilidade da escala em uma segunda

etapa. O uso de duas fases para a etapa da purificação é indicado por estudos como de Gountas, Gountas, Reeves e Moran (2012), pois as etapas são complementares e a validação de face confere a vantagem de se ter um instrumento mais propenso à aprovação pelos testes estatísticos de confiabilidade.

Para melhor relatar tais fatos, esta seção é dividida, na sequência, entre os dois temas: a validação de face e o cálculo da confiabilidade.

Procedimentos para validação de face

A purificação da escala tem objetivo de identificar prematuramente os itens que possuem problemas de redação ou incongruência com o constructo que se pretende mensurar. Para tanto, DeVellis (2003) sugere que parte do processo seja realizada por validação de face, pois a leitura de especialistas é complementar ao uso de ferramentas estatísticas na função de identificar itens deslocados do objeto da escala.

A validação de face observa se os itens da escala parecem claros e adequados aos especialistas para a mensuração (Fink, 1995). Embora haja certa subjetividade no julgamento destes envolvida na validação de face, Gountas *et al*. (2012) afirmam que este teste é importante para a purificação de uma escala, pois pode detectar alguma falha de construção dos itens anteriormente aos cálculos.

O procedimento de validação de face incluiu, inicialmente, o convite a especialistas para a realização de um grupo focal. Durante a realização do procedimento, na tentativa de obter maior objetividade, os especialistas devem ser convidados a preencher um formulário no qual precisam assinalar se cada item apresentado é adequado e/ou claro, o qual é utilizado como base para as análises. Anteriormente a cada preenchimento de formulário, deve-se proceder a uma breve discussão sobre o item, na qual cada participante pode expor a sua opinião e assim sintonizar o grupo em torno do constructo em questão.

A seleção dos sujeitos pode ser feita de forma intencional, desde que estejam dentro do perfil desejado ao estudo, preferencialmente, pessoas que estudem, trabalhem ou convivam próximas ao fenômeno estudado, ou seja, especialistas.

O objetivo desta etapa, como afirmam Anastasi (1988), DeVellis (2003) e Bright, Vine, Wilson, Masters e Mcgrath (2012), é verificar, em um grupo de especialistas, se os itens da escala podem ser considerados adequados à mensuração de um Construto.

Após a definição da validação de face, encadeia-se, pela elaboração do protocolo, a purificação da escala via cálculo da confiabilidade. No caso deste artigo, tal fase é subsidiada pelo cálculo da confiabilidade ancorado na teoria de resposta ao item (TRI) através do modelo nominal.

Confiabilidade através da TRI

Confiabilidade, passo tradicional nos principais protocolos de elaboração de escalas, não é garantia de validade de Construto; é, na verdade, um passo intermediário que auxilia na purificação da escala. Purificar significa preparar a escala para o teste de validação. Provavelmente, o método mais conhecido e utilizado para se estimar a confiabilidade de uma escala é o cálculo do coeficiente alfa proposto por Cronbach (1951).

Embora a estimativa alfa para a confiabilidade seja certamente a mais utilizada, não é imune a críticas de adequação. Para Sijtsma (2009), o cálculo do alfa para confiabilidade interna é mais uma tradição do que uma escolha técnica, pois o cálculo do alfa despreza a variabilidade natural da amostra.

Ainda sobre as críticas ao uso do coeficiente alfa, Maroco e Garcia-Marques (2006) apregoaram que o mesmo instrumento apresenta valores sensivelmente diferentes se aplicados a diferentes amostras. Thompson (2002), por sua vez, afirma que a mesma medida, quando administrada a uma amostra de sujeitos mais homogênea ou mais heterogênea, produz escores de confiabilidade diferentes. Em

situações deste tipo, claramente, para os críticos, o coeficiente alfa não é capaz de mensurar a confiabilidade do instrumento, o que foi mensurado foi a homogeneidade da amostra.

Pelos motivos expostos, também, TenBerge e Socan (2004) afirmam que o cálculo do coeficiente alfa não é uma mensuração de consistência interna, tampouco uma medida de unidimensionalidade. Sijtsma (2009), a seu turno, atesta que, embora haja um entendimento coletivo da academia de que o cálculo do coeficiente alfa seja capaz de mensurar o quanto todos os itens estão mensurando a mesma dimensão, o teste apresenta escores elevados quando aplicado tanto em escalas unidimensionais quanto multidimensionais, ou seja, não contribui efetivamente se o objetivo for garantir que apenas um Construto foi alvo de mensuração. Isso se dá pelo fato das escalas unidimensionais possuírem reconhecidamente maior potencial de mensuração que as multidimensionais (Schjoedt & Shaver, 2012).

Por fim, Pasquali e Primi (2003) afirmam que, em cálculos da teoria clássica dos testes como o coeficiente alfa ou mesmo a correlação item-total corrigido (CITC), há uma incongruência lógica, pois o escore de cada item é testado contra um escore total, que é constituído por todos os itens do teste, inclusive o que está sendo analisado. A partir disso, presume-se que os outros itens já estejam validados *a priori* ou, de outra forma, não faria sentido serem incluídos nos cálculos. Mas, paradoxalmente, se já se soubesse a princípio da confiabilidade dos itens, não haveria sentido em testá-los.

Uma alternativa promissora ao uso do coeficiente alfa é a teoria de resposta ao item (TRI), desenvolvida, pela psicometria, para avaliar testes psicológicos dicotômicos unidimensionais, baseada em uma variável latente, como em Lord (1952). Devido à complexidade dos cálculos, baseados em ogiva normal e função integral, a TRI permaneceu, durante décadas, subutilizada; porém, com o advento do *software* especializado e com a substituição do uso da ogiva normal pela função logística, a técnica se tornou acessível e ganhou mais espaço na academia.

Sua aplicação mais famosa, no Brasil, é na área de educação, para resolver o problema do cálculo da nota em testes que não possuem o mesmo peso em todas as questões. A TRI permite que indivíduos que tenham o mesmo número de acertos possuam escores diferentes, sendo a única forma de igualar os escores em caso de coincidência de resposta em todas as questões (Drasgow, Levine, Tsien, Williams, & Mead, 1995).

A TRI, segundo Lord e Novick (1968), calcula a probabilidade de resposta ao item levando em consideração a característica do item (parâmetros do item) e também da variável latente (Construto). Essa relação probabilística é definida pela curva característica do item (CCI), que, segundo Chernyshenko, Stark, Chan, Drasgow e Williams (2001), é uma função logística da probabilidade de uma resposta ser assinalada.

Conquanto seja utilizada com sucesso na área de educação, como, por exemplo, no cálculo das notas do Exame Nacional do Ensino Médio (Andrade & Klein, 2005), seu uso em administração ainda é muito restrito. Há, contudo, aplicações da TRI para análise de confiabilidade, como em Bernardi, Bussab e Camargo (2009).

Para se entender a TRI é preciso compreender inicialmente que todas as estimativas são sobre o item, e não sobre a amostra, assim como que conceitos como amostragem probabilística são definitivamente secundários. O importante nessa técnica é o comportamento do item, independente do grupo em que esteja sendo testado.

A TRI é um conjunto de modelos matemáticos que procura representar a probabilidade de um indivíduo dar uma resposta certa a um item, como função dos parâmetros do item e da habilidade dos respondentes. Quanto ao seu procedimento, segundo Pasquali e Primi (2003), os dados se destinam à identificação da natureza do item: dicotômico ou não dicotômico; do número de populações envolvidas: uma ou mais; e da quantidade de traços latentes que está sendo mensurada: uma ou mais.

Os modelos selecionáveis se diferenciam, inicialmente, pelo número de parâmetros do teste e pelo tipo da variável. Quanto ao número de parâmetros, eles podem ser de um parâmetro (somente a dificuldade do item), dois parâmetros (a dificuldade e a discriminação) ou três parâmetros (a

discriminação, a dificuldade e a probabilidade de resposta correta dada por indivíduos de baixa habilidade). Já em relação ao tipo de variável, apresenta-se como nominal ou razão.

O modelo de TRI destinado à análise de dados nominais é conhecido também como TRIN e foi proposto, originalmente, por Bock (1972), fazendo uso de variável dicotômica. Assim, adapta-se perfeitamente à análise de confiabilidade de escalas em si do tipo Likert, desde que reduzido a duas opções de atitude: positiva e negativa.

A TRIN é baseada em função logística e em distribuição da curva normal (*sigma*). E, devido ao seu caráter de orientação ao item, não exige qualquer esforço relativo à amostragem, sobretudo aconselhando apenas que esta seja a maior possível, em virtude da necessidade de calibragem (Pasquali & Primi, 2003).

Para o cálculo da confiabilidade, só é necessário o uso do parâmetro discriminação, porém o modelo de três parâmetros é mais completo, rigoroso e proporciona uma melhor estimação, sendo tarefa de o pesquisador ler os resultados de acordo com os objetivos do teste em questão. O importante é a quantidade de informação do item, determinada pelo cálculo da variância, como observado em Bernardi *et al*. (2009). Na prática, esse valor não é dado pelo *software* por não ser específico para cálculo de confiabilidade. Ao invés disso, calcula-se a quantidade de informação para cada ponto da distribuição da variável latente e exibe-se o resultado em forma de gráfico. O cálculo da área do gráfico estima se há informação suficiente ou não para considerar o item confiável, mas, infelizmente, a literatura não apresenta os parâmetros para esta solução, ao contrário, esta se apropria de outra solução que simplifica os cálculos, estimando a confiabilidade com base no valor do parâmetro denominado *a*, que é um dos parâmetros do algoritmo do cálculo.

Aprofundando um pouco o conceito, a TRIN, assim como a TRI original, possui alguns parâmetros que auxiliam a interpretação dos dados. Quais sejam *a* para discriminação do item, *q* para aptidão, também denominado traço latente ou habilidade, e *b* para dificuldade do item.

Quanto às leituras, Bernardi *et al*. (2009) afirmam que a confiabilidade de cada item da escala pode ser observada através do valor do parâmetro de discriminação *a*, que informa a inclinação da curva no momento de inflexão. Os valores assumidos por *a* vão de 0 a 3, o valor nulo para quando não há discriminação e 3 para discriminação perfeita. Quando o objetivo for o cálculo da confiabilidade, o que se busca é a estimação de discriminação do item (Bernardi, Bussab, & Camargo, 2009) e em relação aos valores desejados para o parâmetro *a*, se seu valor for inferior a 0,85 haverá informação suficiente para considerar o item confiável (Thissen, 1986). Há também uma segunda faixa de valores confiáveis quando *a* é superior a 1,7 (Baker, 2001); portanto, pode-se afirmar que o item é confiável se o parâmetro de discriminação não possuir valores entre 0,85 e 1,70. É necessário ressaltar que esta formulação é exclusiva do modelo de três parâmetros.

A principal vantagem da TRI na determinação da confiabilidade de uma escala é que assume a heterogeneidade da contribuição de informação de cada item à mensuração da escala, pois assume-se uma função de informação para cada item (Lord, 1980). Portanto a TRI anula a necessidade do cálculo tradicional de confiabilidade, como a estimativa alfa de Cronbach (Zagorsek, Stough, & Jaklic, 2006).

O resultado desta etapa é uma escala purificada e composta de itens confiáveis para se mensurar o Construto intencionado, cujo próximo passo a mostrar é a validação, a prova final de uma escala e última etapa do protocolo de elaboração, a qual é vista na próxima subseção.

Validação da escala

A proposta deste estudo é elaborar uma escala preditiva, com boa capacidade de previsão. Essa tarefa, contudo, diferencia-se daquelas encontradas em Churchill (1979), Rossiter (2002) e DeVellis (2003), devido ao caráter nominal da escala.

A primeira distinção que deve ser feita é relativa aos termos validação e confiabilidade. Enquanto a validação se refere à capacidade da escala mensurar algum Construto externo, a confiabilidade de uma medida é referente à sua consistência, ou seja, sua capacidade de ser aplicada em situações similares e aferir a mesma medida (Churchill, 1979).

Para Cronbach e Meehl (1955), se a previsão dos Construtos for feita *a priori*, deve ser chamada de validação preditiva e, caso ocorra simultaneamente à realização do estudo, deve ser conhecida como validação concorrente. No entanto prever com antecedência e observar a confirmação são tarefas mais árduas para as quais se necessita maior conhecimento teórico, embora resultem em maior precisão do estudo (Zemack-Rugar, Corus, & Brinberg, 2012).

A condição para que um Construto seja admissível pela ciência é que, pelo menos, alguns dos seus correlatos sejam observáveis (Churchill, 1979; Know *et al*., 2013). No caso da atitude, essa atribuição, de acordo com Allport e Hartman (1925), é dada à expressão verbal ou à expressão observável, que são as reduções empíricas da variável latente atitude.

Percebe-se, então, que a variável não precisa ser diretamente observável, podendo assumir sua forma latente, que pode ser articulada em uma rede de relacionamentos nomológicos válida e capaz de ser predita. Assim, Cronbach e Meehl (1955) propuseram que o investigador deve conhecer a teoria de interesse e, ao fazer isso, identificar quais são os pontos mensuráveis do tema. Somente após realizar este esforço, ele poderá escolher quais Construtos irá testar.

Os relacionamentos nomológicos, como exemplificado por Zemack-Rugar *et al*. (2012), são sistemas fechados de leis que constituem a teoria relacionada ao Construto estudado e às predições feitas sobre o mesmo, as propriedades observáveis dos Construtos envolvidos e os próprios Construtos.

A validação nomológica tem como requisito básico que os Construtos sejam conhecidos previamente. Porém, no caso da mensuração de atitude, esta atividade aparenta ser possível, por se tratar de um Construto já delimitado pela literatura (Anilkumar & Joseph, 2012). É importante frisar que, ao se isolar uma variável, é possível que haja mais de um componente presente, porém dá-se o nome daquilo que se acredita existir em maior quantidade, promovendo a maior responsividade ao Construto isolado.

Com isso em mente, o primeiro passo de uma validação é, então, entender a rede nomológica de relacionamentos. O segundo passo é a validação do próprio Construto, momento em que são apresentadas as regras pelas quais é possível mensurá-lo (determinação do método de pesquisa). O terceiro passo pode ser dado de duas formas, ou o pesquisador faz as predições e estabelece o que espera encontrar nos Construtos (validação preditiva); ou não faz predição alguma e observa como as relações se revelam ao analisar os dados (validação concorrente).

Porém, para Churchill (1979), um Construto deve ser mensurado de duas ou mais formas, pois só assim se poderá saber se os resultados obtidos são válidos. Esta técnica é chamada validação convergente e busca uma alta correlação entre os instrumentos que mensuram o mesmo Construto. Caso as duas escalas mensurem Construtos diferentes, o que deverá se observar é a não correlação entre elas, e esta operação é denominada validação divergente.

A validação é um dos pontos de conflito entre os três principais protocolos vigentes. Enquanto Churchill (1979) e DeVellis (2003) assumem que é possível realizar validações nomológicas que permitam uso de análise fatorial confirmatória, Rossiter (2002) não compartilha deste entendimento e adota a forma preditiva.

A validação preditiva congrega diversas técnicas, sendo a principal o tratamento na matriz multitraço multimétodo (MTMM), como proposto por Campbell e Fiske (1959), que chamaram de traço o Construto a ser testado e método a forma de coleta de dados e abordagem de pesquisa escolhida.

A rigor, segundo levantamento bibliográfico empreendido, o esforço de validação preditiva sobre o Construto atitude deve ter sua origem histórica atrelada ao texto seminal de Grim (1936), o qual apresenta algo similar à MTMM, com apenas um traço, entretanto, sob o nome de validação específica.

Outra variação da MTMM é observável no trabalho de Bauerband e Galupo (2014), pois o desenvolvimento da escala utiliza também um único traço, porém este é levado a campo para teste em dois grupos amostrais diferentes.

Tanto Grim (1936) quanto Bauerband e Galupo (2014) realizaram validações preditivas, o que exige um exaustivo e precioso trabalho de compreensão da literatura para determinar o comportamento dos instrumentos em desenvolvimento antes de se recorrer aos trabalhos de campo. Ambos não realizaram a MTMM completa pelo fato do Construto atitude ser reflexivo e não possuir traço concorrente que justifique a dupla mensuração, ao passo que utilizar mais de um método confere a validação desejada.

A operacionalização da proposta MTMM propõe que sejam escolhidos dois traços (variáveis observáveis) e dois métodos. Cada traço deve ser mensurado por dois métodos diferentes e, na tabulação cruzada, a mesma variável deve ter alta correlação nos dois métodos de coleta de dados (validação convergente), enquanto as variáveis diferentes devem ter baixas correlações (validação discriminante), também, em ambos os métodos. Assim, a correlação entre os mesmos traços, mensurados por diferentes métodos, deve ser mais forte que os dois traços entre si.

A validação convergente é amplamente utilizada de forma isolada, independente do conjunto MTMM. Se, a rigor, para Campbell e Fiske (1959), toda validação de Construto deveria ter ao menos a mensuração de dois traços e dois métodos, algumas variações são possíveis. Existem autores que sugeriram a mensuração de multitraços através de método único, como, por exemplo, Churchill (1979), que sugere uma análise divergente com uso da análise fatorial confirmatória (AFC).

Graças à última sugestão aludida, a estatística de análise fatorial confirmatória está presente em um grande número de estudos em administração, como, por exemplo, em Know *et al.* (2013). Segundo Alina e Caraivan (2012), o pesquisador deve decompor o Construto em subdimensões para facilitar a mensuração e, através da AFC, provar estatisticamente que os relacionamentos teorizados são confirmados empiricamente.

Sendo assim, com o uso da AFC, como proposto por DeVellis (2003), são mensurados os multitraços, mas não em multimétodos, distanciando-se, então, da matriz MTMM. Enfatiza-se, contudo, que, para a mensuração de atitudes, assim como para as demais variáveis reflexivas, a AFC não é indicada pela falta de traços concorrentes observáveis (Diamantopoulos & Siguaw, 2006).

A validação proposta por este protocolo, então, caracteriza-se pela adoção da Matriz Multitraço-Multimétodo de Campbell e Fiske (1959), pela qual é proposto que o mesmo fenômeno, neste caso, a atitude, seja mensurado por dois métodos diferentes, em que um, preferencialmente, deve ser a observação e o outro, obrigatoriamente, o levantamento. Porém, sabendo da característica das variáveis reflexivas, o segundo traço não é necessariamente incorporado à validação preditiva se for mantido o objetivo original de mensurar apenas a atitude.

Tal qual o estudo de Grim (1936), os resultados da observação e do levantamento devem ser comparados. Acredita-se que deve haver concordância mútua de resultados entre a análise dos dados coletados por observação e por levantamento, caso a pesquisa tenha sido corretamente planejada. Variáveis espúrias podem causar interferência nos resultados e, por este motivo, a validação deve observar e levantar dados por questionários de mais de um grupo, seria aconselhável observar ao menos três grupos para cada método.

Finalmente, com o objetivo de apresentar todas as técnicas de coleta de dados sugeridas neste protocolo, é provida, na Tabela 1, a síntese, a qual as descreve e as posiciona a fim de demonstrar suas finalidades e localização dentro do protocolo proposto.

Tabela 1

Técnicas de Coletas de Dados Utilizadas na Pesquisa

Fase/Etapa da pesquisa	Técnica de coleta	Finalidade
Elaboração do Construto	Grupo focal	Definir, junto a especialistas, os limites do Construto e seus atributos.
Elaboração dos itens	Levantamento	Coletar opiniões da população sobre o Construto.
	Grupo focal	Selecionar, junto a especialistas, quais das opiniões coletadas pelo levantamento tornar-se-ão itens da escala.
Purificação da escala	Grupo focal (Validação de face)	Definir, junto a especialistas, por critério de clareza e adequação, se os itens selecionados possuem capacidade de mensuração do Construto.
	Levantamento	Verificar, por meio da TRI, se os itens da escala formarão um grupo confiável de afirmativas para se mensurar o Construto.
Validação	Observação	Mensurar a atitude através da manifestação do comportamento das pessoas.
	Levantamento	Aplicar a escala elaborada e comparar seu resultado com o método da observação em busca de uniformidade de resultados.

Nota. Fonte: Elaboração própria.

Elaborar um instrumento psicométrico válido é uma tarefa árdua que, quando bem-sucedida, fornece um poderoso instrumento à academia para compreender o comportamento de determinada população. Para tanto, o processo de elaboração deve ser rigoroso, pois, dificilmente, os aplicadores do instrumento irão reavaliar as etapas preliminares da construção do instrumento, daí a necessidade do esforço de coleta de dados durante as etapas de desenvolvimento da escala. Porém, o esforço de coleta de dados poderá ser reduzido caso o teste de purificação não exija alterações no instrumento de coleta de dados, ou seja, quando a escala for aprovada na fase de purificação o mesmo banco de dados será utilizado para a fase de validação

O protocolo proposto

Ao final da revisão e discussão teórica, surge a proposta do protocolo de elaboração de escalas para mensuração de atitude. O esquema apresentado na Figura 1 resume o encadeamento lógico das ferramentas escolhidas para o processo.

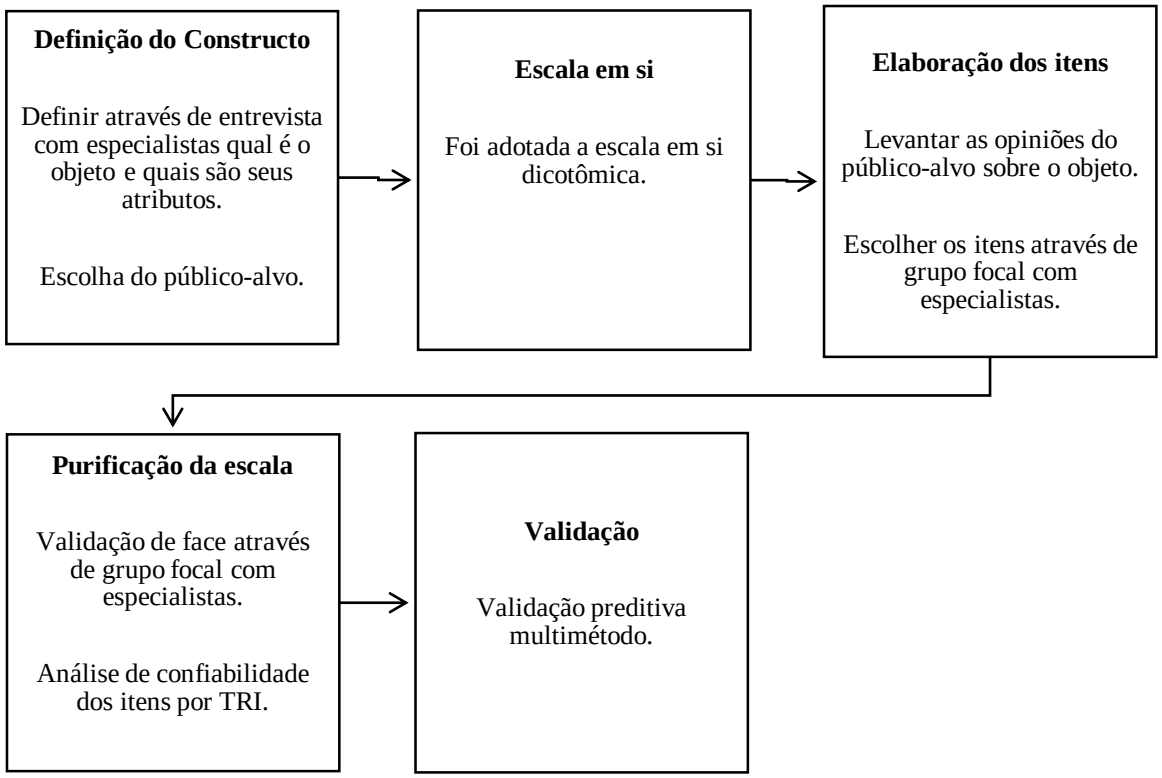

Figura 1. Esquema de Formulação do Protocolo Proposto
Fonte: Elaboração própria.

As inovações no protocolo proposto apresentam-se, então, no uso da TRI, em substituição ao cálculo do coeficiente alfa para estimação da confiabilidade e na retomada da validação preditiva multimétodo (MTMM) de Campbell e Fiske (1959), que, embora quase centenária, estava completamente em desuso em administração para tal finalidade, mas aqui é renovada como alternativa viável e confiável para a validação de escalas reflexivas, como é o caso das escalas de mensuração de atitude.

Além de incorporar novas técnicas ao desenvolvimento de escalas em administração, foi também proposta uma escala em si inovadora a partir da incorporação de melhorias sugeridas pela literatura ao modelo original de Likert (1932), com melhor aderência à mensuração específica de atitude e ao uso da TRI.

Finalmente, este artigo propôs um novo protocolo de elaboração de escalas de mensuração de atitude através de revisão bibliográfica de artigos clássicos e contemporâneos que abordaram o tema por diversas facetas e findou por sugerir um protocolo incremental quanto às etapas e inovador quanto ao conteúdo destas.

Considerações Finais

Com o intuito de atingir o objetivo proposto por este artigo, foi elaborado um protocolo para criação de escalas reflexivas particularmente sensíveis à atitude. Tal variável é aplicável a diversos contextos e exige do protocolo grande versatilidade, pois, embora restrito às variáveis reflexivas, o mesmo pode ser utilizado por pesquisadores de diferentes áreas com distintas finalidades, desde que seu Construto seja compatível com as restrições de aplicação.

Propor uma série de passos lógicos que leve à construção de uma escala não representa por si só uma ideia inovadora, mas a inovação em seus passos para que a proposição como um todo se diferencie do que há disponível na literatura e ganhe espaço como ferramenta legítima e inovadora de construção de escalas. Assim, é importante destacar alguns pontos que enfatizem a contribuição científica deste artigo ao incorporar melhores práticas ao desenvolvimento de escalas de atitude. Como subsídio para tal análise, a Tabela 2 apresenta um resumo das práticas dos principais protocolos vigentes e uma comparação com a proposta atual.

Tabela 2

Comparação entre o Protocolo Proposto e os Protocolos Dominantes

Etapa \ Autor	Churchill (1979)	Rossiter (2002)	DeVellis (2003)	Protocolo proposto
Definição do Construto.	Não abordado.	Questões guias.	Análise subjetiva.	Questões guias.
Escala em si.	Likert.	Likert.	Ao critério do pesquisador.	Dicotômica.
Elaborar itens.	Não abordado.	Combinação subjetiva entre atributos e objetos.	Proposição subjetiva.	Levantados do público-alvo.
Purificação da escala.	Alfa de Cronbach.	Alfa de Cronbach.	Validação de face e Alfa de Cronbach.	Validação de face e TRI.
Validação da escala.	Análise fatorial confirmatória.	Preditiva.	Análise fatorial confirmatória.	Preditiva.

Nota. Fonte: elaboração própria.

A elaboração de qualquer escala inicia com a definição do constructo, tal etapa pode ser considerada a mais importante no processo, pois, sem a definição precisa do que será medido, qualquer mensuração será imprecisa. Neste sentido, entendeu-se que a melhor solução apresentada pela teoria foi de Rossiter (2002), que, ao contrário de Churchill (1979) e DeVellis (2003), evita a subjetividade e alinha tal etapa ao carácter epistemológico das posteriores. Portanto, o protocolo proposto adota a prática das perguntas guias para definir os limites do objeto mensurado.

Na escolha da escala em si, surgiu a primeira inovação, ao contrário de Churchill (1979) e Rossiter (2002), o protocolo proposto não adota o modelo de Likert (1932), tampouco deixa a escolha por conta do pesquisador, como em DeVellis (2003). Para este contexto, foi proposta uma escala mais simples e acurada para mensuração de atitude, que se baseia na definição básica de atitude positiva e negativa. Assim, a escala em si proposta é dicotômica e possui um terceiro campo para anulação da questão, o que se mostrou eficiente em Lake (2014). Tal decisão é particularmente importante para ter melhor aderência ao cálculo da TRI.

Definido o Construto e a escala em si, o próximo passo do protocolo foi a elaboração dos itens. Sob o argumento de manter o alinhamento epistemológico também para esta etapa, não foi admitida subjetividade, contrapondo-se, então, aos protocolos de Rossiter (2002) e DeVellis (2003). Essa atividade atribuiu aos especialistas a tarefa de escreverem os itens que irão compor o questionário, ao invés de sugeri-los por julgamento do pesquisador.

A etapa seguinte, a purificação da escala, envolveu a validação de face, tal qual sugerido por DeVellis (2003), e a coleta de dados no campo, com uma versão preliminar de questionário para cálculo de confiabilidade, apresentando uma segunda inovação, qual seja o uso da técnica TRI para o cálculo ao invés do tradicional uso do alfa de Cronbach, adotado pelos protocolos de Churchill (1979), Rossiter (2002) e DeVellis (2003).

Por fim, a validação se deu de forma preditiva multimétodo nos moldes da MTMM de Campbell e Fiske (1959), em oposição ao uso da análise fatorial confirmatória como em Churchill (1979) e DeVellis (2003), observando o fenômeno por dois métodos concorrentes e estipulando o resultado dos testes. O protocolo proposto se alinha com Rossiter (2002) novamente ao entender que, *a priori*, a confirmação da previsão é também a validação e somente após percorrer todas essas etapas com sucesso é que se pode inferir que o trabalho está encerrado e a escala desenvolvida está apta para o uso.

Referências

Alina, L., & Caraivan, L. (2012) Elaborating a measurement instrument for the flow experience during online information search. *Annals of the University of Oradea, Economic Science Series, 21*(2), 841-847.

Allport, F. H., & Hartman, D. A. (1925). The measurement and motivation of atypical opinion in a certain group. *The American Political Review, 19*(4), 735-760. doi: 10.2307/2939163

Anastasi, A. (1988). *Psychological testing.* New York: Macmillan.

Ander-Egg, E. (1978). *Introducción a las técnicas de investigación social.* Buenos Aires: Nueva Visión.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology, 70*(4), 394-400. doi: 10.1037/h0022280

Andrade, D. F., & Klein, R. (2005). Aspectos quantitativos da análise dos itens da prova do Enem. In Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, *Exame nacional do ensino médio (Enem): fundamentação teórico-metodológica* (pp. 107-112) Brasília: Autor.

Anilkumar, N., & Joseph, J. (2012). Factors influencing the pre-purchase attitude of consumers: a study. *IUP Journal of Management Research, 11*(3), 23-53.

Bagozzi, R. P. (1981). Attitude, intentions, and behavior: a test of some key hypotheses. *Journal of Personality and Social Psychology, 42*(4), 607-627. doi: 10.1037/0022-3514.41.4.607

Baker, F. (2001). *The basics of item response theory* (ERIC Clearinghouse on Assessment and Evaluation). Maryland: College Park MD.

Bardin, L. (2009). *Análise de conteúdo* (4a ed.). Lisboa: Edições 70.

Bauerband, L. A., & Galupo, M. P. (2014). The gender identity reflection and rumination scale: development and psychometric evaluation. *Journal of Counseling & Development, 92*(4), 219–231. doi: 10.1002/j.1556-6676.2014.00151.x

Bearden, W. O., & Netemeyer, R. G. (1999). *Handbook of marketing scales: multi-item measures for marketing and consumer behavior research* (2nd ed.). California: SAGE.

Bernardi, P., Jr., Bussab, W. O. de, & Camargo, R. A. (2009, setembro). Análise da confiabilidade do índice de predisposição para a tecnologia na estrutura da teoria de resposta ao item. *Anais do Encontro Nacional da Associação Nacional de Pós-Graduação e Pesquisa em Administração*, São Paulo, SP, Brasil, 33.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51. doi: 10.1007/BF02291411

Boyd, H. W., Westfall, R., & Stasch, S. F. (1977). *Marketing research. Text and cases* (4th ed.). Illinois: Richard D Irwin Inc.

Bright, E., Vine, S., Wilson, M. R., Masters, R. S., & Mcgrath, J. S. (2012). Face validity, construct validity and training benefits of a virtual reality TURP simulator. *International Journal of Surger, 10*(3), 163-166. doi: 10.1016/j.ijsu.2012.02.012

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. doi: 10.1037/h0046016

Chapa, O., & Stringer, D. (2013). The path of measuring moral courage in the workplace. *SAM Advanced Management Journal, 78*(2), 17-24.

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*(1), 64-73. doi: 10.2307/3150876

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi: 10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. doi: 10.1037/h0040957

Crowther, J. R. (1995). *ELISA. Theory and practice.* Totowa, New Jersey: Springer Science & Business Media.

Derham, P. A. J. (2011). Using preferred, understood or effective scales? How scale presentations effect online survey data collection. *Australasian Journal of Marketing & Social Research, 19*(2), 13-26.

DeVellis, R. F. (2003). *Scale development: theory and applications* (2nd ed.). London: Sage Publications, Inc.

Diamantopoulos, A., & Siguaw, J. (2006). Formative versus reflexive indicators in organizational measure development: a comparison and empirical illustration. *British Journal of Management, 17*(4), 263-282. doi: 10.1111/j.1467-8551.2006.00500.x

Doll, W. J., & Torkzadeh, G. (1991). The measurement of end-user computing satisfaction: theoretical and methodological issues. *MIS Quartely, 15*(1), 5-11. doi:10.2307/249429

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple choice tests. *Applied Psychological Measurement, 19*(2), 143-165. doi: 10.1177/014662169501900203

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155-174. doi: 10.1037//1082-989X.5.2.155

Fink, A. (1995). *How to measure survey reliability and validity.* Thousand Oaks: Sage.

Galton, F. (1880). Statistics of mental imagery. *Mind., 5*(19), 301-318.

Gountas, J., Gountas, S., Reeves, R. A., & Moran, L. (2012). Desire for fame: scale development and association with personal goals and aspirations. *Psychology and Marketing, 29*(9), 680-689. doi: 10.1002/mar.20554

Grim, P. R. (1936). A technique for the measurement of attitudes in the social studies. *Educational Research Bulletin, 15*(4), 95-104.

Guttman, L. (1943, February). A basis for scaling qualitative data. *Annual Meeting of the American Sociological Society*. New York, NY, 38.

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: implications for improving face validity of measures of unobservable constructs. *Journal of Business Research, 57*(2), 98-107. doi: 10.1016/S0148-2963(01)00295-8

Kant, I. (2009). *Crítica da razão pura*. São Paulo: Editora Martin Claret.

Kaptein, M. C., Nass, C., & Markopoulos, P. (2010, April). Powerful and consistent analysis of likert-type ratingscales. *Proceedings of the International Conference on Human Factors in Computing Systems – CHI '10,* New York, NY, USA, 28.

Know, M., Lee, J., Won, W., Park, J., Min, J., Hahn, C., Gu, X., Choi, J., & Kim, D. (2013). Development and validation of a smartphone addiction scale (SAS). *PLoS One, 8*(2), e56936. doi: 10.1371/journal.pone.0056936

Komorita, S. S. (1963). Attitude content, intesity, and the neutral point on a Likert scale. *Journal of Social Psychology, 61*(2), 327-334. doi: 10.1080/00224545.1963.9919489

Lake, C. J. (2014). *Simulating response latitude effects in attitude survey using IRT* (Dissertation). College of Bowling Green State University. Bowling Green, USA.

Lee, S. P., Cornwell, T. B., & Babiak, K. (2012). Developing an instrument to measure the social impact of sport: social capital, collective identities, health literacy, well-being and human capital. *Journal of Sport Management, 27*(1), 24-42.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 1-55.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph, 7). Iowa City, IA: Psychometric Society

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Malhotra, N. K. (2011). *Pesquisa de marketing: uma orientação aplicada* (6a ed.). São Paulo: Bookman.

Mann, P. H. (1970). *Método de investigação sociológica.* Rio de Janeiro: Zahar.

Maroco, J., & Garcia-Marques, T. (2006). Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? *Laboratório de Psicologia, 4*(1), 65-90.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *The Journal of Marketing, 49*(4), 41-50. doi: 10.2307/1251430

Pasquali, L., & Primi, R. (2003). Fundamentos da teoria da resposta ao item. *Avaliação Psicológica, 2*(2), 99-110.

Peabody, D. (1962). Two components in bipolar scales: direction and extremeness. *Psychology Review, 69*(2), 65-73. doi: 10.1037/h0039737

Pérez, A., & Bosque, I. R. del (2013). Measuring CSR image: three studies to develop and to validate a reliable measurement tool. *Journal of Business Ethics, 118*(2), 265-286. doi: 10.1007/s10551-012-1588-8

Petter, S., Rai, A., & Straub, D. (2012). The critical importance of construct measument specification: a response to aguirre-urreta and marakas. *MIS Quartely, 36*(1), 147-155.

Pooja, S., & Sagar, M. (2012). High impact scales in marketing: a mathematical equation for evaluating the impact of popular scales. *Advances in Management, 5*(4), 31-48.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: a meta analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13. doi: 10.1111/j.1745-3992.2005.00006.x

Rossiter, J. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing, 19*(4), 305-335. doi: 10.1016/S0167-8116(02)00097-6

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality Social Psychology Review, 5*(4), 296-320.

Sanches, C., Meireles, M., & Sordi, J. O. de (2011, agosto). Análise qualitativa por meio da lógica paraconsciente: método de interpretação e síntese de informação obtida por escalas likert. *Anais do Encontro de Ensino e Pesquisa em Administração e Contabilidade*, João Pessoa, PB, Brasil, 3.

Schjoedt, L., & Shaver, K. G. (2012). Development and validation of a locus of control scale for the entrepreneurship domain. *Small Business Economics, 39*(3), 713-726. doi: 10.1007/s11187-011-9357-0

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120. doi: 10.1007/s11336-008-9101-0

Sjoberg, G., & Nett, R. (1968). *A methodology for social research*. New York: Harper & Row.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly, 13*(2), 147-169. doi: 10.2307/248922

TenBerge, J. M. F., & Socan, G. (2004). The greatest lower bound to the reliability of a test and hypothesis of unidimensionality. *Psychometrika, 69*(4), 613-625. doi: 10.1007/BF02289858

Thissen, D. (1986). *Multilog: item analysis and scoring with multiple category response models.* Mooresville: Scientific Software.

Thompson, B. (2002). *Contemporary thinking on reliability issues*. Newbury Park: Sage.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529-554. doi: 10.1086/214483

Tomas, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: two factors or method effects. *Structural Equation Modeling, 6*(1), 84-98. doi: 10.1080/10705519909540120

Turner, R., & Zolin, R. (2012). Forecasting success on large projects: developing reliable scales to predict multiple perspectives by multiple stakeholders over multiple time frames. *Project Management Journal, 43*(5), 87-99. doi: 10.1002/pmj.21289

Viswanathan, M., Sudman, S., & Johnson, M. (2004). Maximum versus meaningful discrimination in scale response: implications for validity of measurement of consumer perception about products. *Journal of Business Research, 12*(57), 8-24. doi: 10.1016/S0148-2963(01)00296-X

Zagorsek, H., Stough, S., & Jaklic, M. (2006). Analysis of the reliability of the leadership practices inventory in the item response theory framework. *International Journal of Selection and Assessment, 14*(2), 180–191. doi: 10.1111/j.1468-2389.2006.00343.x

Zemack-Rugar, Y., Corus, C., & Brinberg, D. (2012). The "response-to-failure" scale: predicting behavior following initial self-control failure. *Journal of Marketing Research, 69*(12), 996-1014. doi: 10.1509/jmr.10.0510

Dados dos Autores

Rafael Lucian
Rua Jean Émile Favre, 422, Imbiribeira, 51200-060, Recife, PE, Brasil. E-mail: rlucian@fbv.edu.br

Jairo Simião Dornelas
Avenida dos Economistas, S/N Cidade Universitária, 50670-902, Recife, PE, Brasil. E-mail: jairo@ufpe.br