PMKT
REVISTA BRASILEIRA DE PESQUISAS DE MARKETING, OPINIÃO E MÍDIA

Editor-Chefe: Fauze Najib Mattar | Editor Adjunto: Claudio Alberto de Moraes
Sistema de avaliação: Triple Blind Review
Publicação: Associação Brasileira de Empresas de Pesquisa (ABEP)
Idiomas: Português e Inglês | ISSN: 2317-0123 (on-line)

Uso da teoria de resposta ao item para determinar a confiabilidade de escalas

Use of item response theory to measure the scales reliability

Rafael Lucian*
Faculdade Boa Viagem – DeVry | FBV, Recife, PE, Brasil

RESUMO

Este ensaio teórico dedica-se a estudar por meio de quais procedimentos é possível considerar escalas válidas e aptas para o uso como instrumento científico legítimo. As escalas são ferramentas de mensuração que compõem o instrumentário meio da ciência, a fim de construir conhecimento. O interesse deste artigo, em particular, é sobre o tratamento estatístico para determinar a confiabilidade das escalas. Sendo assim, a proposta aqui é discutir em profundidade se há argumentação sustentável para que a academia adote as técnicas da Teoria de Resposta ao Item (TRI) como estatística segura para o cálculo da confiabilidade de escalas em detrimento de técnicas clássicas como o Coeficiente Alfa de Cronbach. Para tanto, inventariou-se o estado da arte relativo ao tema e revelaram-se as propriedades da TRI ao ponto de ser possível concluir que esta é uma técnica promissora que já possui condições de aplicação imediata nos estudos como ferramenta única e segura para o cálculo da confiabilidade das escalas.

PALAVRAS-CHAVE: Teoria de Resposta ao Item; Confiabilidade das escalas; Coeficiente Alfa de Cronbach.

ABSTRACT

This theoretical essay aims to the study through which procedures it is possible to considered a scale valid as a legitimate scientific instrument. Scales are measurement tools that make up the toolbox through science, in order to build knowledge. The interest of this article, in particular, is on the statistical treatment to determine the reliability of the scales. Therefore, the proposal here is to discuss in depth is no sustainable argument for the health management to adopt the techniques of Item Response Theory (IRT) as safe statistics to calculate the scales of reliability at the expense of classic techniques as the coefficient Cronbach's alpha. Therefore, inventoried up state of the art on the subject and proved properties of TRI to the point it is possible to conclude that this is a promising technique that already has immediate application conditions in administration studies as unique and safe tool for calculating the reliability of the scales.

KEYWORDS: *Item Response Theory; Reliability of scales; Cronbach's Alpha.*

Submissão: 22 ago. 2016
Aprovação: 22 mar. 2017

***Rafael Lucian**

Doutor em Administração pela Universidade Federal de Pernambuco. Coordenador do Núcleo de Apoio à Pesquisa na Faculdade Boa Viagem – DeVry | FBV.
(CEP 51200-060, Recife, PE, Brasil).
E-mail: rlucian@fbv.edu.br
Endereço: Rua Jean Emile Favre, FBV, Ipsep, 51200-060, Recife, PE, Brasil.

1 INTRODUÇÃO

Mensuração, segundo Crowther (1995), é uma técnica que faz uso de instrumentos de precisão para se medir qualidades desejadas com base numérica. Sendo assim, a princípio, qualquer coisa observável 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; ao invés disso, seu objetivo é prover um mecanismo de análise que gere informação e sirva de fomento para uma tomada de decisão inteligente (Pooja & Sagar, 2012).

As escalas, nesta perspectiva, servem ordinariamente aos tomadores de decisões e nesta visão a qualidade da mensuração é, em algum grau, a própria qualidade da decisão (Pooja & Sagar, 2012). Assim, como Sanches, Meireles e Sordi (2011) argumentam, tanto mercado quanto academia buscam mensurações que propiciem melhor qualidade de dados e promover melhoramentos no processo de mensuração é uma contribuição legítima, necessária e atual à academia. Mais objetivamente, Robertson (2012) afirma que prover melhorias sobre os processos de mensuração torna-se uma necessidade à academia. Esse ensaio teórico endossa essa premissa e busca cooperar nessa direção.

No sentido de promover tal contribuição, a mensuração deveria conferir confiabilidade e validade aos dados coletados em campo, porém diversos autores (Thompson, 2002; Ten Berge & Socan, 2004; Maroco & Garcia-Marques, 2006; Vieira & Dalmoro, 2008; Sijtsma, 2009) vêm contestando a capacidade dos testes estatísticos vigentes em gerar escalas com tais características. A inquietação de tais autores se concentra em aspectos básicos da Teoria Clássica dos Testes (TCT), como a variabilidade à amostra que, neste ensaio teórico, serão prioritariamente evidenciados por meio da escrutinação das propriedades matemáticas do Coeficiente Alfa de Cronbach utilizado dominantemente para estabelecer a confiabilidade de escalas.

Alternativamente, influenciados por teóricos de psicologia e educação, pesquisadores (por exemplo, Matteucci, Mignani, & Veldkamp, 2012; Schultz, Salomo & Talke, 2013; Lucian & Dornelas, 2015) admitem considerar o uso da Teoria de Resposta ao Item (TRI) como substituto natural das técnicas atuais de cálculo da confiabilidade, porém, é necessária uma reflexão sobre tal prática, afinal trata-se de uma apropriação.

Neste mister, o objetivo central deste ensaio teórico é discutir em profundidade se há argumentação sustentável para que a academia adote as técnicas da TRI como estatística segura para o cálculo da confiabilidade de escalas em detrimento do tradicional Coeficiente Alfa de Cronbach.

2 CÁLCULO DA CONFIABILIDADE

O uso de escalas de mensuração exige certos cuidados metodológicos, pois seus resultados se tornam importantes indicadores para tomadas de decisões. Alguns instrumentos de avaliação de performance ou Psicometria (como por exemplo satisfação, lealdade ou atitude) são utilizados diariamente por corporações globais que depositam fé em sua capacidade de ler corretamente a verdade. Pesquisadores contemporâneos vêm desenvolvendo e testando modelos esquemáticos e teorias a partir de observações empíricas baseadas em coleta de dados por levantamento e, em última instância, em mensuração por meio de escalas.

Entretanto, para se alinhar aos pressupostos do paradigma científico dominante tais instrumentos devem ser consistentes o suficiente para mensurarem o construto entendido de forma precisa, ou ao menos, de forma constante. Sendo assim, a capacidade de uma escala mensurar igualmente o mesmo nível de interação por meio de duas ou mais intervenções de campo distintas, é chamada de confiabilidade. Embora bastante particular, a confiabilidade se apoia em alguns pressupostos da Teoria Clássica dos Testes (TCT) que se querem revelados.

Na Teoria Clássica dos Testes, o escore total (T) de uma pessoa é composto de duas partes, seu escore verdadeiro (V) mais o erro de mensuração (E). Tal erro é uma variável que pode assumir valores positivos ou negativos, permitindo que o T seja maior ou menor que o V.

Para a confiabilidade de escalas, então, o que se procura é a razão entre a variância de V e T, porém tal indicador não é simplesmente calculado pelo fato de não se poder conhecer o valor de V e também pela teorização de que o escore verdadeiro deveria ter pouquíssima ou nenhuma variação entre os testes, resultando em variância zero.

À parte disso, a elaboração de novas escalas normalmente envolve um intenso processo de aproximação e erro. Dada a regra, os pesquisadores sugerem à novas escalas, itens capazes de mensurar determinado construto e verificam tal capacidade em campo por meio da aplicação de testes estatísticos de confiabilidade. Por vezes, esses procedimentos de pré-teste podem indicar que o instrumento de mensuração não é suficientemente confiável e deve ser modificado.

Outro cenário é a elaboração de testes de confiabilidade da escala para confirmar que não houve graves problemas de amostragem (ligado ao trauma permanente da amostragem não probabilística, ou pior, por conveniência). Tal aplicação torna-se uma regra prática optativa, mas recorrente, pois há uma singela preocupação contraditória entre adotar escalas confiáveis e submetê-las novamente ao teste de confiabilidade. Diga-se que uma verificação de não confiabilidade pós-teste sugere que a validação pré-teste foi indevida (descartando-se, a propósito, erros de amostragem que prejudicariam qualquer mensuração por si só).

Para tal, provavelmente, o método mais conhecido e utilizado para se estimar a confiabilidade de uma escala é o cálculo do Coeficiente Alfa de Cronbach (1951). Tal cálculo é inspirado em um processo, já em desuso, conhecido como teste-reteste, que tinha o objetivo de testar a mesma escala duas ou (de preferência) mais vezes com a mesma amostra ao longo do tempo.

Uma técnica anterior ao Teste Alfa, que certamente a inspirou e ainda pode ser encontrada nos principais *softwares* estatísticos, é a de corte-ao-meio ou *Split-Half* (CAM), que consiste em elaborar duas escalas equivalentes no mesmo questionário. Teoricamente, a pessoa deveria dar a mesma resposta ao responder a mesma coisa duas vezes por métodos diferentes, desde que as escalas tivessem consistência interna. Essa consistência as validaria conjuntamente.

As escalas validadas pelo método corte-ao-meio deveriam ter todos os seus itens no mínimo duplicados (escritos de forma diferente, mas equivalentes) e isso limitava a capacidade de exploração, já que o limite para o tamanho do questionário sempre foi a tolerância do respondente. Insurgindo-se contra essa técnica, Cronbach (1951) afirmou que, para obter uma melhor interpretação, a escala não deve ser divisível em pequenos blocos menores: deve ser única, sem subescalas e sem escalas duplicadas; isso, ainda segundo o autor citado, auxiliaria na validação interna da escala.

A contribuição de Cronbach (1951), entretanto, foi elaborar um cálculo mais simples, no qual os itens são correlacionados entre si internamente e não em cruzamentos entre as duas escalas equivalentes do CAM. Assim, os pesquisadores não mais necessitariam criar itens de verificação, já que o Alfa leva em consideração o item pela média da escala. O novo processo tornou a pesquisa por escalas mais simples, rápida e eficiente.

3 A TRADIÇÃO DO ALFA DE CRONBACH

O objetivo dos testes de confiabilidade é simples: medir quanto o indivíduo é consistente em suas respostas; entretanto, tal mensuração sempre exigiu dupla coleta de dados, quais sejam em momentos distintos ou duplicando as escalas no mesmo questionário.

Cronbach (1951) foi, então, revolucionário a seu tempo ao propor uma forma de cálculo de confiabilidade em que se exige coleta única de dados, poupando tempo, esforço e garantindo maior agilidade às pesquisas empíricas, motivo pelo qual seu Coeficiente Alfa é, até os dias atuais, o método mais utilizado para se purificar uma escala.

O Coeficiente Alfa é o indicador de consistência interna e varia naturalmente entre 0 e 1, por ser uma razão entre o escore verdadeiro e o total, entretanto ele pode assumir valores negativos em algumas situações não convencionais nas quais há correlação negativa entre os itens da escala.

A atribuição de escores abaixo de zero para o coeficiente pode ser fruto de uma inversão em parte dos itens durante a coleta de dados que não foi devidamente corrigida na tabulação final, ou caso a dita escala esteja na verdade mensurando construtos diferentes entre seus itens (Henson, 2001). Na

prática, entretanto, não se admite valores negativos, ou seja, caso os escores sejam próximos ou abaixo de zero a única interpretação desejável é que não há confiabilidade.

Como regra prática, admite-se que valores acima de 0,6 (Malhotra, 2013) ou 0,7 (Gouveia, Santos, & Milfont, 2013) são sarrafos mínimos para se conferir a escala a título de confiável. Porém, Streiner (2003) enfatiza que altos valores no Alfa de Cronbach devem ser lidos como falha na variabilidade de mensuração da escala. Seria como se todos os itens fossem um só, então, para se dizer confiável, o escore máximo aceitável deve ser 0,9. Finalmente, em termos de leitura de resultados, admite-se escalas confiáveis se estas possuírem valores de Alfa de Cronbach entre 0,6 e 0,9.

Embora a estimativa Alfa para a confiabilidade seja certamente a mais utilizada, não é imune às críticas de adequação. Para Sijtsma (2009), o cálculo do Alfa para confiabilidade interna é mais uma tradição do que uma escolha técnica. Isso se revela para aquele autor, quando o mesmo observa a vasta literatura que critica severamente o uso desta técnica para os fins de estimativa da confiabilidade, como por exemplo, Thompson (1994), Vacha-Haase (1998), Wilkinson (1999) e Maroco e Garcia-Marques (2006).

Um dos principais argumentos destes autores é que o cálculo do Alfa despreza a variabilidade natural da amostra. Maroco e Garcia-Marques (2006) apregoaram que o mesmo instrumento apresenta valores sensivelmente diferentes se aplicados a diferentes amostras. Thompson (2002) afirma que a mesma medida quando administrada a uma amostra de sujeitos mais homogêneos ou mais heterogêneos, produz escores de confiabilidade diferentes. Em situações deste tipo, o Coeficiente Alfa não é capaz de mensurar claramente a confiabilidade do instrumento, o que foi mensurado foi a homogeneidade da amostra.

Nesta mesma perspectiva, para teóricos como Caruso (2000), Yin e Fan (2000) e Streiner (2003) o Alfa de Cronbach não é capaz de mensurar a confiabilidade da escala, pois seu resultado é muito dependente das características da amostragem, como já comentado. Caruso (2000) e Henson, Kogan e Vacha-Haase (2001) enfatizam que, quanto mais heterogênea for a amostra, maior será a variação do escore total e, consequentemente, maior será o valor do coeficiente de confiabilidade.

Nesta perspectiva, os escores obtidos pelo Alfa de Cronbach só poderão ser estendidos à escala de mensuração caso a homogeneidade da amostra seja também revelada.

Além disto, denota-se que há uma alta sensibilidade do Teste Alfa ao número de casos. Duhacheck e Iacobucci (2004) afirmam que, quanto menor o tamanho da amostra, maior será o valor da estimativa de confiabilidade. Sendo assim, Maroco e Garcia-Marques (2006) afirmam que os valores relativos ao Alfa devem sempre ser interpretados à luz das características da medida a que se associa e da população na qual essa medida foi feita, fato esse já reconhecido anteriormente pelo próprio Cronbach (1951) em sua publicação seminal.

Pelos motivos expostos, também Ten Berge e Socan (2004) afirmam que o cálculo do Coeficiente Alfa não é uma mensuração de consistência interna, tampouco uma medida de unidimensionalidade. O Coeficiente Alfa é forte dependente do comprimento da escala. Cortina (1993) testou e confirmou que mensurações baseadas em intervalos com muitos pontos elevam os valores do Alfa de Cronbach, mesmo que tal variação não influencie em nada na consistência interna e no escore verdadeiro.

Sijtsma (2009) 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 pode apresentar 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.

Finalmente, Pasquali e Primi (2003) afirmam que, em cálculos da teórica clássica dos testes como o Coeficiente Alfa, 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. Isto sugere que os outros itens são adequados ou de outra forma não faria sentido serem incluídos nos cálculos. Mas se já se soubesse, a princípio, da confiabilidade dos itens, não haveria sentido em testá-los.

Em relação a todas as limitações supracitadas, existe a alternativa da Teoria de Resposta ao Item, que é não apenas outra forma de mensurar a confiabilidade, mas uma alternativa à própria TCT.

4 ALTERNATIVA DA TRI

Uma evolução promissora ao uso do Coeficiente Alfa é a Teoria de Resposta ao Item que foi 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 sugestão de Birnbaum (1968) de se substituir a 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. Testes em larga escala, como o Exame Nacional do Ensino Médio (ENEM), são construídos por diversos professores que, provavelmente, não conseguirão elaborar questões com o mesmo grau de dificuldade. Entretanto, seria igualmente improvável que, propositalmente, qualquer teste conseguisse repetir a mesma dificuldade em suas diversas versões, ou, mais improvável ainda, mensurar adequadamente o peso de cada questão correta de acordo com sua dificuldade.

Pela TCT, se uma questão for assinalada corretamente por 80% dos candidatos pode-se dizer que sua dificuldade é de 0,8. Embora o cálculo seja simples ele depende da habilidade dos respondentes, ou seja, se por exemplo, o mesmo teste for aplicado a um grupo de alunos com menor habilidade, o percentual de acertos da mesma questão poderá baixar para 50% e assim, a dificuldade seria de 0,8 para o primeiro grupo e de 0,5 para o segundo grupo. Desta forma, a TRI surge como a solução para se mensurar a dificuldade da questão independentemente da amostra.

A Teoria de Resposta ao Item permite que indivíduos que tenham o mesmo número de acertos possuam escores diferentes caso a habilidade dos candidatos seja igualmente diferente, afinal seu objetivo não é contar o número de respostas corretas e sim mensurar a habilidade do respondente. Em verdade, a única forma de igualar os escores é em caso de coincidência de resposta em todas as questões (Drasgow, Levine, Tsien, Williams, & Mead, 1995).

A TRI, então, segundo Lord e Novick (1968), calcula a probabilidade de resposta ao item levando em consideração a característica do item (parâmetros do item) e também a habilidade em relação a variável latente (construto). Essa relação probabilística é definida pela curva característica do item (CCI) que, segundo Chernyshenko, Stark, Chan, Drasgow, & Williams (2001) é uma função logística da probabilidade de uma resposta ser assinalada. Nos extremos da CCI verifica-se que um indivíduo com habilidade igual a 3,0 terá aproximadamente 100% de probabilidade de acerto enquanto outro com escore -3,0 praticamente não possui chances de responder corretamente a questão.

Embora 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 pesquisas acadêmicas ainda é muito restrito. Há, contudo aplicações da TRI para análise de confiabilidade, como em Lucian e Dornelas (2015).

Para se entender a TRI é preciso entender inicialmente que todas as estimativas são sobre o item e não sobre a amostra e 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. Sendo assim, as análises do item são feitas com base na variável latente que independe do comportamento da amostra específica (Matteucci, Mignani, & Veldkamp, 2012).

Denomina-se Teoria de Resposta ao Item a um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma resposta certa a um item, como função dos parâmetros do item e da habilidade dos respondentes.

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

(dificuldade, discriminação e chance de acerto ao acaso). Já em relação ao tipo de variável, apresenta-se como nominal ou razão (Lucian & Dornelas, 2015).

A dificuldade do item é representada pela letra *b* e pode variar entre -4,0 e +4,0 e resultados próximos de zero indicam dificuldade média. Nesse ponto, é importante enfatizar que qualquer um dos três modelos é capaz de mensurar a habilidade (Θ) que é expressa na mesma escala da dificuldade do item, ou seja é representada pelo eixo *x* no gráfico da CCI.

O segundo parâmetro, presente apenas no modelo de dois ou três parâmetros é a discriminação do item representado pela letra *a* e é o escore mais importante para o objetivo de mensurar confiabilidade das escalas. Este escore tem sua variabilidade usual entre 0 e +3,0 sendo que, quanto maior mais reativo, ou seja possui maior capacidade de detectar pequenas variações na habilidade dos respondentes. Existe a possibilidade de *a* retornar valores negativos caso a probabilidade de acerto e Θ sejam inversamente proporcionais, caso que indica inversão dos itens na escala ou problemas de formulação na questão. Na CCI o parâmetro *a* é calculado no ponto de inflexão (Bacci, 2012).

O terceiro parâmetro é representado pela letra *c* e indica a chance de acerto ao acaso, ou seja, *c* busca o número de acertos aferidos por pessoas de habilidade muito baixa. Não parâmetros absolutos para a chance de acerto ao acaso, deve-se observar seu valor em relação ao cálculo 1/*n*, sendo que *n* é o número de alternativas. Caso os escores de *c* distanciem-se muito do esperado isto indica problemas na construção da questão.

Para o cálculo da confiabilidade, o interesse particular é pelo modelo de dois parâmetros, pois apenas o escore *a* é levado em consideração atualmente. A TRI, entretanto, é uma técnica muito recente e, com o avanço do interesse dos pesquisadores por seus benefícios, outros parâmetros poderão ter seus benefícios revelados.

Existe ainda outro modelo particular de teoria de resposta ao item que pode ser adotado nos estudos que é conhecida como TRIN. Proposto originalmente por Bock (1972), a TRIN faz uso de variáveis dicotômicas e está associada ao tratamento de escalas nominais. Assim, adapta-se perfeitamente à análise de confiabilidade de escalas em si do tipo Likert, em que há 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 a amostragem (Pasquali & Primi, 2003). O importante é mensurar a quantidade de informação do item, como observado em Bernardi, Bussab e Camargo (2009).

Na prática, esse valor não é dado diretamente pelos *softwares*, pois a TRI não é específica 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, ou seja, estima-se a confiabilidade com base no valor do parâmetro *a*.

Os valores assumidos por *a*, como já mencionado, vão de 0 a +3,0 (podendo assumir valores negativos nas situações especificadas anteriormente), o valor nulo para quando não há discriminação e 3 para discriminação perfeita. Quando maior o valor de *a*, significa que maior é a probabilidade de sucesso dos respondentes que possuem maior Θ que, neste caso, representa a presença da variável latente.

Quando o objetivo for o cálculo da confiabilidade, o que se busca é a estimação de discriminação do item 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. Há também uma segunda faixa de valores confiáveis quando *a* é superior a 1,7; sendo assim, pode-se afirmar que o item é confiável se o parâmetro de discriminação não possuir valores entre 0,85 e 1,70 (Lucian & Dornelas, 2015).

A composição de tais valores em duas faixas não é usual em TCT e, a princípio, pode confundir o usuário da TRI, entretanto explica-se que uma escala é dita confiável quando seus itens possuem a capacidade de discriminar a presença da variável latente na amostra, sendo assim, os valores intermediários possuem menor representatividade que os extremos.

Clarificando ainda mais o conceito, exemplifica-se que, em uma escala hipotética para mensuração da interação do usuário com sistemas inteligentes, espera-se que a escala discrimine os indivíduos

que possuem muito pouca interação ou os que possuem realmente muita interação, assim se saberá que o corte foi rigoroso e os resultados são confiáveis.

A principal vantagem da TRI na determinação da confiabilidade de uma escala é que ela 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). Sendo assim, a TRI anula a necessidade do cálculo tradicional de confiabilidade, como a estimativa Alfa de Cronbach (Zagorsek, Stough, & Jaklic, 2006).

As exigências para uso da TRI em sentido amplo são de que os itens sejam unidimensionais (cada escala mensurado apenas um construto), tenham características de eventos independentes (baixa correlação interna entre os itens) e sejam monotônicos (probabilidades de acerto proporcionais entre os itens e em um só sentido).

Devido à característica de sobreposição das probabilidades em escalas do tipo Likert (não é monotônica, pois as repostas podem flutuar livremente entre os extremos), o cálculo do parâmetro *c* não é adequado (Gutierrez, 2005), porém a discriminação do item indicado pelo coeficiente *a* é independente à distribuição de probabilidade e atende aos propósitos do cálculo da confiabilidade (Bernardi, Bussab, & Camargo, 2009), logo sugere-se que seja adotado o modelo de dois parâmetros para tal finalidade.

5 CONSIDERAÇÕES FINAIS

Diante dos benefícios apresentados pela Teoria de Resposta ao Item sobre a Teoria Clássica dos Testes apresentada, em particular em seu uso para o cálculo da confiabilidade de escalas em detrimento do Coeficiente Alfa de Cronbach, aconselha-se a adoção da TRI como instrumento estatístico para purificação de escalas e teste de confiabilidade.

Respondendo finalmente ao objetivo do estudo, afirma-se que a Teoria de Resposta ao Item é uma técnica promissora que já possui condições de aplicação imediata como ferramenta única e segura para o cálculo da confiabilidade das escalas. Sendo assim, acredita-se que sua adoção para tal finalidade é uma evolução natural em relação ao Coeficiente Alfa de Cronbach tal qual este foi em relação à técnica CAM.

A principal resistência, contudo, à popularização da TRI é certamente a complexidade dos cálculos, incluindo a ainda falta de **intuitividade** dos *softwares* específicos. Existem algumas opões que podem ser adquiridas para sistemas operacionais Windows como o ConQuest 3, Facets, RUMM2030, WINMIRA, Winsteps e Xcalibre além da opção para Mac Quest.

Alguns oferecem versões para estudo ou são completamente gratuitos como o Bigsteps, ConstructMap, Facets-DOS, Ganz Rasch, ICL, jMetrik, Minifac, MULTIRA e WinLLTM. Tais programas, entretanto, não são intuitivos e alguns chegam a exigir que o usuário insira linhas de código para realizar os cálculos. Até onde se pode constatar, nenhum deles é capaz de importar arquivos de outras planilhas eletrônicas, sendo obrigatório a conversão para TXT ou CSV.

Considerando a usabilidade, as ferramentas de importação e exportação e a diversidade de modelos incluídos (unidimensional, multidimensional, escalar, nominal, dicotômico ou politômicos), o melhor programa para cálculo da Teoria de Resposta ao Item parece ser o IRTPRO que possui versões pagas e versões gratuitas para estudo com algumas limitações.

Recomenda-se, então, que os futuros interessados em tratar dados e purificar escalas de mensuração façam uso da TRI como instrumento de cálculo da confiabilidade dos instrumentos. O benefício dessa prática irá transcender as práticas de pesquisa e colaborar com resultados mais precisos que, em última instância, irão beneficiar os profissionais e executivos em suas práticas gerenciais que dependem do avanço das pesquisas na área.

REFERÊNCIAS

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*, Brasília: O Instituto.

Bacci, S. (2012). Longitudinal data: Different approaches in the context of item-response theory models. *Journal of Applied Statistics, 39*(9), 2047-2065. doi:10.1186/2196-0739-2-1

Bernardi, P. Júnior, Bussab, W. de O., & Camargo, R. A. (2009). Análise da Confiabilidade do Índice de Predisposição para a Tecnologia na Estrutura da Teoria de Resposta ao Item. *Anais do XXXIII Encontro da ANPAD.* São Paulo: ANPAD.

Bimbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.). *Statistical theories of mental test scores* (pp. 397-472), Reading, MA: Addison-Wesley.

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

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60*(2), 236-254. doi: 10.1177/00131640021970484

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. doi: 10.1207/S15327906MBR3604_03

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104.

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

Crowther, J. (1995) *Oxford advanced learner's dictionary*. Oxford: Oxford University Press.

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

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89*(5), 792-808.

Dutra, L. H. de A. (2010). *Introdução à epistemologia.* São Paulo: Editora UNESP.

Gouveia, V. V., Santos, W. S., & Milfont, T. L. (2013). O uso da estatística na avaliação psicológica: Comentários e considerações práticas. In C. S. Hurtz (Org). *Avanços e polêmicas em avaliação psicológica*. São Paulo: Casapsi Livraria e Editora Ltda.

Gutierrez, G. C. (2005). *Estimação das escalas dos construtos capital social, capital cultural e capital econômico e análise do efeito escola nos dados do Peru-PISA 2000* (Dissertação de Mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, 200p. Rio de Janeiro.

Henson, R. K. (2001). Undestanding internal consistency reliability estimates: A conceptual primer on coeficiente alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177-189.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. A. (2001). Reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement*, *61*(3), 404-420. doi: 10.1177/00131640121971284

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.

Lucian, R., & Dornelas, J. S. (2015), Mensuração de atitude: Proposição de um protocolo de elaboração de escalas. *RAC. Revista de Administração Contemporânea (Online)*, *19*(1), 157-177.

Malhotra, N. (2013). *Review of marketing research*. Emerald Group Publishing Limited. Bingley.

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.

Matteucci, M., Mignani, S., & Veldkamp, B. P. (2012). The use of predicted values for item parameters in item response theory models: An application in intelligence tests. *Journal of Applied Statistics*, *39*(12), 2665-2683. doi: 10.6092/unibo/amsacta/3241

Pasquali, L., & Primi, R. (2003). *Fundamentos da Teoria de Resposta ao Item – TRI*. Avaliação Psicológica, 2(2), 99-110.

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.

Robertson, J. (2012). Likert-type scales, statistical methods, and effect sizes. *Communications of the ACM*, *55*(5), 6-7. doi: 10.1145/2160718.2160721

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.

Schultz, C., Salomo, S., & Talke, K. (2013). Measuring new product portfolio innovativeness: How differences in scale width and evaluator perspectives affect its relationship with performance. *Journal of Product Innovation Management*. *30*(2), 93-109. doi: 10.1111/jpim.12073

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

Streiner, D. L. (2003). Starting at the beginning: An introduction to Coefficient Alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99-103. doi: 10.1207/S15327752JPA8001_18

Ten Berge, 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

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

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*(1), 837-847.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6-20. doi: 10.1177/0013164498058001002

Vieira, K. M., & Dalmoro, M. (2008). Dilemas na construção de escalas tipo Likert: O número de itens e a disposição influenciam nos resultados? *Anais do XXXII Enanpad*. Rio de Janeiro.

Wilkinson, L. (1999). Task force on statistical inference, APA Board of Scientific Affairs. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604.

Yin, P., & Fan, X. (2000). Assessing the reliability of beck depression inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*(2), 201-223. doi: 10.1177/00131640021970466

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