ISSNe: 2182.2883 | ISSNp: 0874.0283
Disponível em: http://dx.doi.org/10.12707/RIV15003

ARTIGO TEÓRICO/ENSAIO

Contributo para a Discussão da Avaliação da Fiabilidade de um Instrumento de Medição

Contributions to the Discussion on the Assessment of the Reliability of a Measurement Instrument

Contribución a la Discusión sobre la Evaluación de la Fiabilidad de un Instrumento de Medición

Fernanda Daniel*; Alexandre Gomes da Silva**; Pedro Lopes Ferreira***

Resumo

Enquadramento: A fiabilidade de um determinado instrumento reporta-se à consistência dos resultados obtidos aquando da sua administração. A inspeção da fiabilidade de um instrumento assume, nos dias de hoje, caráter quase imperativo na apresentação dos dados empíricos. Dependendo do propósito da apresentação e do conceito medido, o estudo da fiabilidade pode incluir vários procedimentos.
Objetivos: Mapear as principais técnicas de medição da fiabilidade e os seus algoritmos.
Principais tópicos em análise: Para a medição da estabilidade temporal apresentamos o *teste-reteste* e as formas equivalentes. Na medição da consistência interna são apresentados os modelos alfa de Cronbach, Kuder-Richardson e da bipartição de Spearman-Brown.
Conclusão: Dado que o estudo de fiabilidade de uma escala está diretamente relacionado com o que se pretende medir ou comparar é impossível postular receitas padrão para o uso dos referidos estimadores. Por esse facto propomos uma ponderação cuidada na escolha do mais adequado ao estudo em causa.

Palavras-chave: Reprodutibilidade dos testes; questionários; psicometria

Abstract

Theoretical framework: The reliability of an instrument refers to the consistency of the results obtained during its administration. Nowadays, reliability assessment is almost imperative in the presentation of empirical data. Depending on the purpose of the presentation and the concept measured, the study of reliability can include multiple procedures.
Objectives: To map the main techniques for measuring reliability and their algorithms.
Main topics of analysis: The test-retest and equivalent forms are used to measure temporal stability. Cronbach's alpha, Kuder-Richardson and Spearman-Brown split-half models are used for measuring internal consistency.
Conclusion: Since the study of reliability of a scale is directly related to what we intend to measure or compare, it is impossible to postulate standards for the use of the above-mentioned estimators. For that reason, we propose a careful consideration in choosing the most suitable estimator for the study to be developed.

Keywords: Reproducibility of the tests; questionnaires; psychometrics

Resumen

Marco contextual: La fiabilidad de un instrumento se refiere a la consistencia de los resultados obtenidos durante su administración. La inspección de la fiabilidad de un instrumento conlleva, en estos días, un carácter casi imperativo en la presentación de los datos empíricos. Dependiendo de la finalidad de la presentación y del concepto medido, el estudio de la fiabilidad puede incluir varios procedimientos.
Objetivos: Mapear las principales técnicas de medición de la fiabilidad y sus algoritmos.
Principales temas de análisis: Para la medición de la estabilidad temporal presentamos el test-retest y las formas equivalentes. En la medición de la consistencia interna se presentan los modelos alfa de Cronbach, Kuder-Richardson y la bipartición de Spearman-Brown.
Conclusión: Dado que el estudio de la fiabilidad de una escala está directamente relacionado con lo que se pretende medir o comparar, es imposible postular recetas estándar para el uso de los estimadores. Por eso, proponemos una ponderación cuidada en la elección del más adecuado para el estudio en cuestión.

Palabras clave: Reproducibilidad de resultados; cuestionarios; psicometría

* Ph.D., Professor Auxiliar, Instituto Superior Miguel Torga e Centro de Estudos e Investigação em Saúde da Universidade de Coimbra, 3000-370, Coimbra [fernanda-daniel@ismt.pt]. Contribuição no artigo: Conceção, redação e aprovação da versão final
** Ph.D., Professor Coordenador, Coimbra Business School - ISCAC/IPC, 3040-316, Coimbra, Portugal [alexmfgs@gmail.com]. Contribuição no artigo: Redação e revisão dos algoritmos. Morada para correspondência: Coimbra Business School - ISCAC/IPC, Quinta Agrícola – Bencanta, 3040-316, Coimbra, Portugal.
*** Ph.D., Professor Associado, Centro de Estudos e Investigação em Saúde da Universidade de Coimbra, 3000-370, Coimbra [pedrof@fe.uc.pt]. Contribuição no artigo: Revisão da versão final

Recebido para publicação em: 07.01.15
Aceite para publicação em: 16.10.15

Introdução

Medir significa estabelecer uma concordância entre a forma de medir e o que está a ser medido. Este processo é relativamente fácil de ser operacionalizado quando estamos em presença de variáveis diretamente observáveis. Contudo, quando as variáveis são latentes (não observadas nem medidas diretamente), esse processo torna-se mais difícil, como é o caso da medição da ansiedade, da satisfação ou da qualidade de vida através de escalas com vários itens.

Atualmente é grande a variedade de instrumentos de medição, que podem assumir a designação de testes, escalas ou inventários "de acordo com a relação que se assume existir entre os itens, ou questões, que compõem o questionário" (Ribeiro, 1999, p. 76). Em cada ano cresce o número de instrumentos de medida elaborados em diferentes países, o que obriga à sua validação e à realização de testes de fiabilidade. Neste artigo focar-nos-emos na fiabilidade.

Até há relativamente pouco tempo, a inspeção da fiabilidade era um processo moroso e difícil de realizar, se tivermos por referência os curtos prazos estabelecidos para as investigações académicas contemporâneas. Com o incremento da informática e devido à maior facilidade de acesso a programas estatísticos, estes processos tornaram-se cada vez mais fáceis de realizar pois é agora possível realizar cálculos complexos em frações de tempo, impensáveis há umas décadas atrás.

Um instrumento de medição deve ser inequivocamente fiável. A fiabilidade dos instrumentos de medição depende, em última instância, das boas características dos seus itens. Para se obter um instrumento fiável poder-se-á ter que equacionar eliminar, substituir ou rever os itens.

O termo *fiabilidade* pretende significar a qualidade daquilo ou de quem é fiável. No contexto da inquirição o termo apresenta especificidades que importa refletir. Fiabilidade reporta-se à consistência de resultados obtidos pelos mesmos indivíduos, quando inquiridos em diferentes momentos ou num determinado momento, sendo neste último caso a fiabilidade determinada a partir de itens equivalentes (Anastasi, 1977; Freeman, 1980). Na linguagem científica encontramos uma panóplia terminológica para nomear o substantivo que aqui designamos por fiabilidade (do inglês, *reliability*). Alguns clássicos nesta área apresentam nas suas traduções para língua portuguesa os termos garantia (Freeman, 1980), precisão (Anastasi, 1977) ou mesmo fidelidade (Bryman & Cramer, 1993). Em autores que escrevem em língua portuguesa os termos fiabilidade (Hill & Hill, 2012; Maroco & Garcia-Marques, 2006) e fidedignidade (Vaz-Serra, Ponciano, & Freitas, 1980) são os utilizados.

Estimar a fiabilidade de um teste é um procedimento fundamental quando recorremos a instrumentos de medição que utilizam escalas de vários itens. Na década de noventa do século passado e no início do século XXI, estimar a fiabilidade de um teste ou de uma escala ou de um inventário tornou-se um procedimento rápido, resultado do incremento de pacotes estatísticos que, de forma célere em associação com plataformas cada vez mais interativas e amigáveis, permitiram a utilização massiva destes procedimentos.

A fiabilidade pressupõe também a reprodutibilidade de resultados. Anne Anastasi, cientista cujo brilhantismo é plasmado numa vasta obra onde a psicometria é desconstruída, afirma que a precisão se refere "à consistência de resultados obtidos pelos mesmos indivíduos em diferentes oportunidades ou com diferentes conjuntos de itens equivalentes" (1977, p. 84). No mesmo sentido, Freeman afirma que a garantia se reflete na consistência dos resultados efetuados em várias avaliações, significando "o grau em que os resultados obtidos estão isentos daqueles defeitos internos suscetíveis de provocar erros de medição inerentes aos próprios itens e à estandardização" (1980, pp. 73-74).

Desenvolvimento

Existem várias estimativas da fiabilidade de um teste, podendo "haver tantas variedades de precisão do teste quantas as condições que influem nos seus resultados, pois qualquer uma dessas condições pode não ter significado para determinado objetivo e, assim, ser classificada como variância do erro identificadas por cada uma" (Anastasi, 1977, p. 85). Dependendo da forma de estimar a fiabilidade, utilizam-se vários procedimentos. O pressuposto subjacente ao estudo da fiabilidade é calcular o tamanho do erro. Assim, no caso da inexistência de variabilidade nos resultados, não haveria erro e a fiabilidade seria igual a 1. De notar que a definição de erro é complexa porque está

associada ao conceito mensurado. Se o teste mede a depressão, a diferença encontrada nos resultados entre as duas administrações (variância do erro) pode ser atribuída tanto a flutuações aleatórias (e neste caso falamos de erro aleatório), como pode igualmente ser devida a um diferente estado emocional (e neste caso não podemos afirmar que se trata de erro aleatório). Se objetivarmos medir a estabilidade temporal de um teste temos que nos assegurar que não existem diferenças no estado emocional entre as duas aplicações. Este procedimento pode ser controlado através da inclusão de perguntas sobre experiências intermediárias significativas entre as administrações.

Estabilidade temporal

Diz-se que um instrumento apresenta estabilidade temporal se os seus resultados se mantiverem constantes ao longo do tempo, ou seja, "[r]ealizando-se duas vezes o mesmo teste, a correlação entre os resultados dá-nos uma indicação acerca da estabilidade dos resultados no tempo" (Laveault & Grégoire, 2002, p. 150). Os procedimentos para avaliar a estabilidade temporal ou fiabilidade externa implicam a administração de duas versões do mesmo teste (teste-reteste) ou de duas versões de testes equivalentes.

Teste-reteste

Uma forma de estimarmos a estabilidade temporal de um conjunto de perguntas ou itens equivalentes pressupõe a utilização de um teste-reteste. Para estimar esta estabilidade temporal é necessário administrar o conjunto de perguntas ou itens equivalentes a um grupo e correlacionar a administração dos valores obtidos com os de outra administração efetuada num outro momento. Se o teste for fiável é expectável que, apesar do tempo decorrido entre as duas administrações e se nenhuma alteração significativa ocorrer entretanto, as respostas registadas para cada indivíduo se mantenham inalteradas ou quase.

O conceito subjacente à fiabilidade teste-reteste pode ser extrapolado para a nossa vida quotidiana e para medidas físicas. Imagine-se então que se fizeram duas pesagens consecutivas de uma mesma pessoa, uma às 12 horas e uma outra às 13 horas. O que pensar se entre as duas medições se verificar um aumento de peso de um quilograma? Uma das hipóteses que poderemos colocar para explicar este aumento será o desequilíbrio da balança. Posicionemo-nos agora no papel de um investigador que administrou duas versões de um teste e imaginemos que a sua pontuação duplicou entre as duas administrações. O que pensar? No caso da balança, após a surpresa inicial, provavelmente se levantará a hipótese de ingestão excessiva de alimentos. O mesmo se passa com o resultado obtido no nosso conjunto de perguntas ou itens equivalentes: Ou o teste não é fiável ou entre as duas administrações aconteceu algo que explicará a diferença nos resultados.

As diferenças encontradas nos testes podem assim ser atribuídas a diferentes fatores: (i) administrador, (ii) administrado, ou (iii) cenário. Relativamente ao *administrador* imaginemos que este apresentou, numa das aplicações do teste, um comportamento percecionado pelo respondente como inapropriado. Será que este comportamento se refletirá negativamente nos resultados? Poderá contribuir para um preenchimento pouco rigoroso, pouco profissional ou mesmo desonesto do teste, da escala ou do inventário, logo interferindo nos resultados? Por outro lado, imaginemos que o administrador demonstrou excessiva simpatia a par de uma narrativa longa e inapropriada sobre o instrumento de medição. Poderão estes comportamentos ativar no inquirido respostas de acordo com as normas sociais? É consabido que "o desejo de aceitação social pode enviesar parâmetros avaliados em investigações científicas, constituindo uma ameaça à sua validade, pelo que deve ser controlado" (Poínhos et al., 2008, p. 223).

No tocante ao *administrado* imagine uma de múltiplas possibilidades: No espaço que mediou as duas administrações alguns dos inquiridos ficaram desempregados. Esta nova situação de inatividade forçada é vivida por alguns dos respondentes com elevados níveis de ansiedade. Se o conceito a medir for a ansiedade-estado (estado emocional transitório ou condição do organismo humano caracterizado por sentimentos desagradáveis de tensão e apreensão conscientemente percebidos, consentâneos com o aumento de atividade do sistema nervoso autónomo; Spielberger, 1983), as diferenças encontradas não deverão ser atribuídas a uma baixa fiabilidade teste-reteste mas sim a alterações nos níveis de ansiedade-estado em resultado da nova situação vivida. Neste caso estaríamos em presença de um teste que, apesar de apresentar uma aparente baixa fiabilidade teste-reteste, é fiável. É pertinente mencionar que a

fiabilidade dos instrumentos de medição que avaliam estados de humor podem demonstrar elevada oscilação, isto é, é provável verificarem-se diferenças significativas entre as pontuações. Contrariamente, os instrumentos de medição que avaliam características de personalidade devem apresentar pontuações estáveis ao longo do tempo.

Por último, relativamente ao *cenário*, imagine que a sala onde está a ser administrado o conjunto de perguntas ou itens equivalentes está mal isolada e o frio, ou mesmo o barulho do exterior, provocam mal-estar. É expectável que também nesta situação os resultados reflitam o cenário ambiental vivido.

Importa aqui referir que quando utilizamos um instrumento de medição devemos sempre recorrer ao seu manual ou artigo original ou, caso não exista, aos respetivos autores para obtermos indicação sobre o intervalo que deve mediar as duas administrações. Esta informação é preciosa porque a duração temporal que medeia as administrações poderá criar diferentes efeitos relacionados com o conceito a ser medido. Intervalos curtos podem provocar efeitos relacionados com a memória. Pelo contrário, longos intervalos podem possibilitar a aquisição de novos conhecimentos. Anastasi sobre este assunto refere que:

> Podem ser citados facilmente exemplos de testes que apresentem alta precisão em períodos de poucos dias ou semana, mas cujos resultados revelam uma perda quase completa de correspondência quando o intervalo se estende a até dez ou quinze anos. Por exemplo, muitos testes de inteligência para crianças em idades pré-escolares fornecem medidas moderadamente estáveis no período pré-escolar, mas são virtualmente inúteis para predizer o QI na idade escolar ou adulta. (1977, p. 32)

A idade dos sujeitos tem igualmente importância crucial. Anastasi sugere que o período que medeia um teste-reteste deve ser menor em crianças, porque as mudanças desenvolvimentais nas crianças são discerníveis em períodos curtos (1 mês ou menos) e independentemente da idade o intervalo teste-reteste não deve exceder 6 meses (1977). O mesmo pode suceder quando estamos perante testes de fiabilidade em condições de saúde nas quais se esperam progressões muito diferentes.

Além disto, convém aqui também reportar que a interpretação do resultado da fiabilidade teste-reteste deve ter em consideração o facto de um teste se seguir ao outro. A experiência da primeira administração pode afetar o desempenho do sujeito na segunda administração.

Em nenhum teste, mesmo que meça um traço da personalidade estável, se deve esperar que demonstre uma fiabilidade teste-reteste perfeita com correlação igual a 1, pois existem inúmeros fatores que influenciam as pontuações. Em síntese, a fadiga, os níveis diferentes de concentração e de motivação, as diferentes condições do meio ambiente (a temperatura, os barulhos, as distrações ambientais), os efeitos da prática e da aprendizagem, o intervalo entre as aplicações, os acontecimentos pessoais inesperados no decorrer deste intervalo e os erros inerentes à administração são alguns das possíveis variações que interferem com a variância do erro.

Quando um instrumento de medição é operacionalizado numa escala compósita de itens múltiplos (ex.: através de soma ou médias dos seus itens) pode utilizar-se o coeficiente de correlação linear de Pearson para obter uma estimativa da fiabilidade entre as formas. Em baixo poderá observar uma das fórmulas utilizadas no cálculo deste coeficiente de correlação *r*:

$$r = \frac{n \times \sum XiYi - \sum Xi \times \sum Yi}{\sqrt{\left[n \times \sum Xi^2 - \left(\sum Xi\right)^2\right] \times \left[n \times \sum Yi^2 - \left(\sum Yi\right)^2\right]}}$$

onde x e y representam as variáveis.

Por exemplo: Imagine-se que se administrou um instrumento de medição composto por quatro itens que avalia a rotina laboral dos trabalhadores de uma fábrica de lanifícios. Passado um tempo voltou a administrar-se o mesmo instrumento à mesma amostra. Apresentamos em seguida a Tabela 1 que exemplifica o procedimento que deve ser realizado para calcular a estabilidade temporal através do coeficiente de correlação linear de Pearson.

Tabela 1
Procedimento para o cálculo do coeficiente de correlação de Pearson

N.º	Rotina (1.º Momento)					Rotina (2.º Momento)							
	It. 1	It. 2	It. 3	It. 4	X_i	It. 1	It. 2	It. 3	It. 4	Y_i	X_i^2	Y_i^2	X_i*Y_i
1	4	2	4	5	15	4	3	4	5	16	225	256	240
2	2	1	2	2	7	2	2	2	2	8	49	64	56
3	3	2	4	5	14	3	2	4	5	14	196	196	196
4	2	1	3	3	9	2	1	3	4	10	81	100	90
5	2	2	2	3	9	2	2	2	3	9	81	81	81
6	2	1	2	1	6	2	2	2	2	8	36	64	48
7	3	4	2	4	13	3	4	2	4	13	169	169	169
8	2	1	3	1	7	2	2	3	1	8	49	64	56
9	2	1	3	2	8	2	2	3	1	8	64	64	64
10	2	1	2	1	6	1	2	2	1	6	36	36	36
∑	24	16	27	27	94	23	22	27	28	100	986	1094	1036

Nota. N.º = Número atribuído ao inquirido; it. = item; ∑= somatório; X_i e Y_i = variáveis compósitas (X_i = momento 1 e Y_i = momento 2); X_i^2 e Y_i^2 = variáveis compósitas elevadas ao quadrado; X_i*Y_i = produtos das variáveis compósitas

Um valor de correlação de Pearson de 0,978

$$r = \frac{10\times1036 - 94\times100}{\sqrt{\left[10\times986 - (94)^2\right]\times\left[10\times1094 - (100)^2\right]}} = 0{,}978$$

é considerado significativo ao nível de 0,01, querendo com isto significar que não existe probabilidade maior do que uma em 100 de que a correlação na população seja nula. Deve notar-se que a dimensão da amostra e a sua variabilidade pode afetar todas as medidas apresentadas, isto é, um valor de correlação pode ser considerado significativo numa amostra grande e não o ser numa amostra bem mais pequena. Por fim, poderemos afirmar que em pesquisas onde se recolhem dados com testes já aferidos para a população portuguesa este tipo de estimativa não se revela, na nossa ótica, um procedimento fundamental.

Formas equivalentes

Segundo Anastasi, o coeficiente de fiabilidade "é uma medida tanto da estabilidade temporal como da coerência de resposta a diferentes formas de itens (ou formas de teste)" (1977, p. 95).

O método denominado por formas equivalentes, formas alternativas ou paralelas é semelhante ao método teste-reteste na medida em que são feitas duas administrações do mesmo teste. Existem, contudo, diferenças assinaláveis entre estes dois métodos. Enquanto no método teste-reteste são administradas as mesmas versões nas duas sessões, nas formas equivalentes as versões equivalem-se, ou seja, as versões devem ser iguais nas instruções, na forma e em todas as demais características, e similares no conteúdo. Pretende-se, com este método, eliminar dois tipos de vieses encontrados no método teste-reteste: O facto de os indivíduos poderem recordar-se do teste anterior e os possíveis efeitos da prática. Este método pretende, assim, eliminar os efeitos da prática e da memória ao testar indivíduos através da utilização de versões comparáveis mas não iguais nas duas sessões.

Alta fiabilidade em formas equivalentes sugere que os itens das duas versões do teste são representativos de uma mesma população de itens que hipoteticamente representam o conceito que está a ser medido. Baixa fiabilidade nas formas equivalentes sugere que os dois formatos do teste não estão a medir a mesma coisa. Convém realçar que a fiabilidade através de um teste de formato equivalente contém muitas das mesmas limitações da fiabilidade do teste-reteste.

Aquando da mensuração da fiabilidade teste-reteste e das formas equivalentes é obrigatório mencionar o intervalo de tempo que mediou entre as duas administrações. Se as duas formas são aplicadas uma a seguir à outra, em sucessão imediata, a correlação mostra apenas a fiabilidade entre as formas (conteúdo) e não entre as ocasiões (estabilidade temporal). Pelas características acima apresentadas percebe-se que esta forma não seja muito utilizada em investigação.

Consistência interna

No caso da consistência interna, designada também por fiabilidade interna por Bryman e Cramer (1993), apenas se aplica uma versão e uma única vez. Imagine-se que se quer avaliar uma variável latente a partir de múltiplos itens. Como é expectável, o grupo de itens deve operacionalizar a variável latente e não uma outra variável. Será que os itens têm consistência interna? Só é possível afirmar que um instrumento de medida tem consistência interna se todos os seus itens contribuírem para a medição da mesma característica. O procedimento da consistência interna para estimar a fiabilidade é hoje um dos métodos mais utilizados na investigação, na medida em que é um método que, além de ser económico (Polit & Hungler, 1992) por requer apenas uma só prova, é também o melhor método para avaliar uma das fontes mais importantes de erros de medição que é a seleção dos itens do teste. O coeficiente alfa de Cronbach é comumente utilizado para estimar a fiabilidade de instrumentos nos quais os itens apresentam múltiplas respostas. A regra básica aqui também é de que os valores se devem situar entre 0,8 e 1,0 (Bryman & Cramer, 1993). Quando um conceito e a sua medição compreendem várias dimensões, é habitual calcularem-se os coeficientes de fiabilidade para cada uma das dimensões subjacentes em vez de calcular um só para a medida no seu todo. A fórmula para calcular o coeficiente alfa é a seguinte:

$$\alpha = \frac{k}{k-1}\left[1 - \sum_{i=1}^{k} \frac{s_i^2}{s_t^2}\right]$$

onde K é número de itens da escala, s_i^2 é variância dos resultados do teste no item I e s_t^2 é variância do teste.

O processo de medição da consistência interna através do coeficiente alfa de Cronbach produz baixas estimativas da fiabilidade do teste, mas sobrestima a fiabilidade de testes de velocidade. Consequentemente, os procedimentos de consistência interna são considerados inapropriados para determinar a fiabilidade de testes de velocidade (Ary, Jacobs, & Razavieh, 1990).

Apesar de ser consensual que uma escala deva ser fidedigna, os valores a partir dos quais se infere da fiabilidade da escala parecem não o ser. Senão vejamos a Tabela 2 produzida por Peterson (1994).

Tabela 2

Valores propostos por vários autores sobre o nível recomendado do α de Cronbach α

Autor	Situação	Níveis recomendados
Davis (1994)	Preditor individual	Acima de 0,75
	Previsão para grupos de 25-50	0,5
	Previsão para grupos acima de 50	Abaixo de 0,5
Kaplan e Sacuzzo (1982)	Investigação fundamental	0,7-0,8
	Investigação aplicada	0,95
Murphy e Davidsholder (1988)	Nível inaceitável	Abaixo de 0,6
	Nível baixo	0,7
	Nível moderado a elevado	0,8-0,9
	Nível Elevado	0,9
Nunnally (1978)	Investigação preliminar	0,7
	Investigação fundamental	0,8
	Investigação aplicada	0,9-0,95

Nota. Adaptado de "A meta-analysis of Cronbach's coefficient alpha", de R. A. Peterson, 1994, *Journal of Consumer Research*, 21(2), p. 382.

Este coeficiente, por apresentar algumas fraquezas, tem levantado algumas críticas tendo autores como Maroco e Garcia-Marques (2006) apresentado alternativas. A fiabilidade compósita definida por Fornell e Larcker (1981) para um fator j com *k* itens é dada por $fc = \frac{(\sum_{i=1}^{k} p_i)^2}{(\sum_{i=1}^{k} p_i)^2 + \sum_{i=1}^{k} e_i}$ onde *p* representa os pesos fatoriais de cada item e *e* representa o erro.

Enquanto o coeficiente alfa de Cronbach é utilizado

em testes com itens com múltiplas respostas, a fórmula *KR20* destina-se a testes em que os itens oferecem apenas duas hipóteses alternativas, como é o caso de verdadeiro/falso, sim/não, ou certo/errado:

$$R_{kR-20} = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i q_i}{\sigma_X^2}\right]$$

em que *K* é o numero de itens, *p* a proporção de respostas corretas, *q* a de incorretas e a variância.

Metade-Metade

Um outro procedimento utilizado para medir se um conjunto de questões ou itens equivalentes apresentam fiabilidade designa-se como metade--metade, do inglês *split-half*, também designada como fiabilidade das metades ou da bipartição. Pode chegar-se a um valor de fiabilidade a partir de uma única administração. Neste caso correlacionam-se as duas metades de um teste. A divisão das metades deve ser baseada em critérios que terão em conta tanto o conceito como a forma como este se encontra operacionalizado. Uma variante deste processo é conhecida como o método par-ímpar, que é talvez o método mais antigo para estimar a consistência interna (Polit & Hungler, 1992).

A fiabilidade metade-metade é semelhante à forma alternativa, dividindo uma única escala em duas. Este método avalia o grau de consistência entre os itens, determinando a consistência interna da escala. Não mede a estabilidade temporal, mas oferece a vantagem de permitir obter uma medida de fiabilidade a partir de uma única administração e assume que todos os itens contribuem de igual forma para a mensuração de um conceito central (Anastasi, 1977; Freeman, 1980).

Imagine-se em presença de uma escala composta por 10 itens. Uma das formas de dividirmos a escala seria decompô-la segundo a numeração que lhe foi atribuída inicialmente, ou seja, uma das metades corresponde aos itens pares e a outra é constituída pelos itens ímpares. Outra forma é dividir os itens em duas grandes metades, ou seja, os cinco primeiros itens com os cinco itens da segunda metade. Importa referir que a opção pela divisão deve estar relacionada com o tipo de teste. A divisão por duas grandes metades, a primeira metade com a segunda metade, pode provocar muitos inconvenientes especialmente em testes organizados com um grau crescente de dificuldade. As metades podem ainda ser criadas através de uma seleção aleatória.

Qualquer coeficiente de correlação obtido através da técnica metade-metade tende a gerar uma estimação sistematicamente inferior ao da escala na sua totalidade, ou seja, a correlação obtida é relativa a metade do teste. As escalas com maiores números de itens geram maiores valores de fiabilidade (Polit & Hungler, 1992). Com o objetivo de superar esta dificuldade foi criada uma fórmula para ajustar o coeficiente de correlação para toda a escala. Obtém--se, assim, um coeficiente que pode ser interpretado da mesma forma que o coeficiente de correlação de Pearson, na medida que varia de 0 a 1. Idealmente, ele deve ser maior ou igual a 0,8 (Bryman & Cramer, 1993) e é obtido através da fórmula seguinte:

$$r = \frac{\frac{1}{2}\left(s_p^2 - s_{p1}^2 - s_{p2}^2\right)}{s_{P1} s_{P2}}$$

onde s_p, s_{p1} e s_{p2} representam respetivamente os valores dos desvios padrão para todos os itens e para cada uma das metades.

Por outro lado, conhecendo o valor de correlação entre as duas metades, é possível estimar o coeficiente de correlação para toda a escala. A equação de correção é denominada de Spearman-Brown e traduz-se no seguinte algoritmo:

$$r_{cSB} = \frac{2r_{xy}}{1 + r_{xy}}$$

Nesta fórmula, representa a correção introduzida por Spearman-Brown e a correlação entre as duas metades. Por exemplo, se o coeficiente de correlação entre as duas metades do teste fosse de 0,65 a estimativa para o conjunto da escala seria a seguinte:

$$r_{cSB} = \frac{2 \times 0{,}65}{1 + 0{,}65} = 0{,}79$$

Podemos ainda, neste contexto, estimar os efeitos na fiabilidade de uma determinada escala aquando do aumento ou do decréscimo do número de itens através da seguinte fórmula:

$$\frac{nr}{1 + (n-1)r}$$

onde *n* é a proporção do número de itens em cada forma e *r* a correlação entre as duas metades. Assim, se o número de itens de uma escala for 60, se se obtiver o valor de *r* igual a 0,5 e se estivermos interessados no aumento do número de itens para 150, a proporção do aumento será de 2,5 (150/60).

Através desta fórmula a estimativa da fiabilidade passará a ser de 0,71 em vez de 0,50:

$$\frac{2,5 \times 0,5}{1 + (2,5 - 1) \times 0,5} = 0,71$$

Caso as duas partes não sejam equilibradas utiliza-se o seguinte algoritmo:

$$ULD = \frac{-r^2 + \sqrt{r^4 + 4r^2(1 - Rr^2)k_1k_2/k^2}}{2(1 - R^2)k_1k_2/k^2}$$

onde r é a correlação obtida acima e k é o número de itens de cada parte (UL *Unequal Length*) (IBM® SPSS®, 2011).

O facto de existirem muitas formas de dividir os itens em dois grupos faz com que se possa também obter múltiplas estimativas de fidedignidade. Por esse facto normalmente utilizam-se apenas o coeficiente alfa de Cronbach e a fórmula de Kuder-Richardson pois traduzem a média de todos as bipartições possíveis.

Análise de sensibilidade dos itens da escala

Existem vários procedimentos que poderão ser realizados para avaliar a sensibilidade dos itens de uma escala. Referimo-nos à análise da matriz de correlação inter-item, ao valor do alfa de Cronbach se cada um dos itens fosse eliminado e à correlação item-total.

A matriz de correlações apresenta a correlação de cada item com todos os outros. Na diagonal da matriz deve encontrar o valor 1, já que a correlação de um item com ele próprio será sempre igual à unidade. Os valores das restantes correlações devem ser elevados e positivos, indicando que os itens medem a mesma variável latente. Se existirem valores negativos deve verificar se os itens em causa estavam na mesma direção conceptual dos outros itens e, se for esse o caso, deve proceder-se à sua recodificação.

É sempre possível verificar, como análise de sensibilidade, o impacto de cada item no modelo global com a medição do alfa sem esse item ou com a correlação desse item com a soma dos restantes. Caso não haja impacto na escala, é preferível a mais simples. Duas medidas que mostram isso são o alfa item-total e a correlação item-total como se pode ver nas fórmulas seguintes.

A análise do valor do alfa de Cronbach quando um determinado item é eliminado permite-nos analisar o impacto da exclusão de um determinado item. É dado pela fórmula

$$\bar{A}_i = \frac{k-1}{k-2}\left(1 - \sum_{\substack{l=1 \\ l \neq i}}^{k} S_l^2/\tilde{S}_i^2\right)$$

em que k é o numero de itens, s_l^2 é variância dos resultados do teste no item I e $\tilde{s}_i^2$ é variância do teste sem o item.

Se da análise da eliminação de determinados itens da escala se verificarem valores de alfa de Cronbach superiores ao valor de alfa da escala total deve considerar a eliminação desses itens da escala. Importa mencionar que valores do alfa de Cronbach inferiores a 0,7 não são considerados favoráveis daí que a possibilidade de eliminação do item deva ser equacionada (Pallant, 2007). Note que só é aconselhável a eliminação de itens de uma escala quando esta está a ser construída ou quando se pretende reduzir o número dos seus itens. Caso contrário, o procedimento de eliminação inibe, em rigor, qualquer comparação dos nossos resultados com a escala original.

Uma outra forma de obtermos informações acerca da consistência interna de uma escala é através da correlação item-total que nos indica o grau segundo o qual cada item se correlaciona com a pontuação total. Esta correlação é dada por $R_i = \frac{cov(X_i,P)-S_i^2}{S_i\tilde{S}_i}$. Caso seja necessário eliminar itens, a correlação item-total permite-nos obter informações sobre os itens que apresentam correlações mais baixas com o resto da escala e, por consequência, sobre os itens a eliminar. Teremos assim uma escala com itens com maior consistência interna apesar de a escala no seu conjunto, caso se tenham eliminado muitos itens, diminuir em termos de fiabilidade medida através da consistência interna. Segundo Pallant (2007) valores baixos (menores que 0,3) indicam que o item não está a medir a escala como um todo, podendo existir alternativa à escala apresentada. Se o alfa de Cronbach da escala global for baixo, por exemplo inferior a 0,7, podem existir itens incorretos, podendo haver necessidade da sua remoção de acordo com baixas correlações item-total.

Conclusão

Todos os instrumentos de medição apresentam vantagens e desvantagens pois dependem de vários fatores como, por exemplo o tipo de dados (escala), a forma como foi operacionalizado o constructo (unifatorial ou multifatorial), o objetivo do instrumento (avaliação de um traço de personalidade ou conhecimentos) e os respetivos itens. Estes fatores

ditarão qual ou quais as medidas para cada caso. Estamos perante uma situação de medida relativa, dado que a fiabilidade ou a consistência está sempre diretamente relacionada com o que se pretende medir ou comparar.

Todos os estimadores apresentarão resultados diferentes para a mesma situação. Geralmente o teste-reteste apresenta valores mais baixos pois depende de mais do que uma avaliação. Por outro lado, algum cuidado deve ser tido no desenho experimental, dado que a análise é feita normalmente em estudos quasi-experimentais não havendo muitas vezes aleatoriedade e equilíbrio nas medições.

Assim, não podemos postular uma receita padrão para o uso dos referidos estimadores, antes propomos uma ponderação cuidada na escolha do mais adequado ao estudo em causa.

Referências Bibliográficas

Anastasi, A. (1977). *Testes psicológicos* (2.ª ed.). São Paulo, Brasil: Editora Pedagógica e Universitária.

Ary, D., Jacobs, L. C., & Razavieh, A. (1990). *Introduction to research in education* (4th ed.). Fort Worth, USA: Ted Buchholz.

Bryman, A., & Cramer, D. (1993). *Análise de dados em ciências sociais: Introdução às técnicas utilizando o SPSS* (2.ª ed.). Oeiras, Portugal: Celta.

Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *1*, 39–50.

Freeman, F. (1980). *Teoria e prática dos testes psicológicos* (2.ª ed.). Lisboa, Portugal: Fundação Calouste Gulbenkien.

Hill, M., & Hill, A. (2012). *Investigação por questionário* (2.ª ed.). Lisboa, Portugal: Edições Sílabo.

IBM® SPSS®. (2011). *IBM SPSS statistics 20 algorithms*. Recuperado de ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf

Laveault, D., & Grégoire, J. (2002). *Introdução às teorias dos testes em ciências humanas*. Porto: Porto Editora.

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.

Pallant, J. (2007). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows* (3th ed.). New York, USA: Mc Graw Hill.

Peterson, R. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, *21*(2), 381–391.

Poínhos, R., Correia, F., Faneca, M., Ferreira, J., Gonçalves, C., Pinhão, S., & Medina, J. L. (2008). Desejabilidade social e barreiras ao cumprimento da terapêutica dietética em mulheres com excesso de peso. *Acta Médica Portuguesa*, *21*, 221–228.

Polit, D., & Hungler, B. (1992). *Investigacion cientifica en ciencias de la salud* (3.ª ed.). México: Nueva Editorial Interamericana.

Ribeiro, J. L. P. (1999). *Investigação e avaliação em psicologia da saúde*. Lisboa, Portugal: Climepsi.

Vaz-Serra, A., Ponciano, E., & Freitas, F. (1980). Resultado da aplicação do Eysenk personality inventory a uma amostra da população portuguesa. *Psiquiatria Clínica*, *21*, 127–132.