As propriedades psicométricas dos instrumentos de hetero-avaliação

Objetivo: Este artigo pretende sistematizar para as propriedades métricas dos instrumentos de hétero-avaliação. **Método:** Para o efeito foi realizada uma revisão da literatura, recorrendo à consulta de manuais e à pesquisa em bases de dados internacionais. **Resultados:** São apresentadas as propriedades métricas, especificamente a reprodutibilidade, validade e responsividade e os seus parâmetros. **Conclusão:** Os instrumentos de heteroavaliação devem ser precisos, válidos e responsivos, de modo a garantir resultados fidedignos.

Palavras-chave: Psicometria, Validade, Reprodutibilidade.

Aim: This article intends to systematize to psychometric properties of the instruments of hetero-evaluation. **Method:** A review of the literature was done, using manuals and international data bases. **Results:** Psychometric properties, specifically reliability, validity and responsiveness and their parameters are presented. **Conclusion:** Hetero-evaluation instruments must be accurate, valid and responsive in order to ensure reliable results.

Keywords: Psychometrics, Validity, Reproducibility.

INTRODUÇÃO

A utilização da versão original de instrumentos de hetero-avaliação pode conduzir a enviesamentos nos resultados, se não for feito um processo rigoroso de tradução e a adaptação cultural. Neste sentido, recomenda-se a realização de testes que avaliem as propriedades de medida (propriedades métricas), nomeadamente, consistência interna, reprodutibilidade, validade e responsividade (Puga, Lopes & Costa, 2012).

O processo de tradução e adaptação é fundamental para testar se há equivalência com a versão original, resolvendo as diferenças de costumes, língua e perceção da saúde entre países e culturas diferentes. Além disso é necessário testar as propriedades de medida, mesmo que já tenham sido testadas no instrumento original, pois pode haver diferenças culturais entre populações

> A compreensão destes conceitos ajudam os enfermeiros e investigadores a conhecerem as vantagens e desvantagens da aplicação de um determinado instrumento/escala para medir um atributo ou função.

distintas, e também permite avaliar se o instrumento adaptado retém as propriedades de medida da versão original (Puga, Lopes & Costa, 2012).

A Fidedignidade, reprodutibilidade e precisão são termos utilizados para avaliar uma importante propriedade psicométrica de instrumentos de avaliação de constructos subjetivos, que é a fiabilidade da medida (Fabrício-Wehbe et al, 2013). A fiabilidade é definida como a consistência com que o instrumento mede o atributo. Esta indica a reprodutibilidade de uma medida, isto é, a capacidade de reproduzir o mesmo resultado quando se aplica o instrumento repetidamente em sujeitos em que não houve alteração do seu estado. A validade de um instrumento de medida consiste no grau que ele mede o que é suposto medir. A Responsividade é considerada habitualmente como a sensibilidade para mudanças, ou seja, é a capacidade que o instrumento tem de medir mudanças pequenas, mas clinicamente importantes que o sujeito desenvolve como resposta a uma intervenção terapêutica efetiva. (Oliveira & Santos, 2011; Polit, Beck & Hungle, 2011).

A seleção do instrumento de avaliação de uma função a ser utilizado tem de ser feita de forma cuidada. Este instrumento (ou escala) tem de ser considerado como uma medida válida da função a ser testada, com uma fiabilidade adequada e que os resultados devem ser suficientemente sensíveis para traduzir alterações clínicas significativas (Hoeman, 2000). Neste sentido, um instrumento clinicamente útil deve ser apropriado para avaliar a função/atributo, ser breve, facilmente aplicável e pouco oneroso (Cavaco & Alouche, 2010).

O objetivo deste artigo é esclarecer os enfermeiros e investigadores sobre as propriedades métricas de um instrumento de hétero-avaliação. A compreensão destes conceitos ajudam os enfermeiros e investigadores a conhecerem as vantagens e desvantagens da aplicação de um determinado instrumento/escala para medir um atributo ou função.

PROPRIEDADES MÉTRICAS

As propriedades métricas (propriedades da medida, psicométricas, clinicométricas) de um instrumento de hétero-avaliação são a Reprodutibilidade, Validade e Responsividade (ou sensibilidade à mudança).

Reprodutibilidade

A reprodutibilidade avalia o quanto um instrumento está livre do erro aleatório ou fornece um resultado reprodutível (Barbetta & Assis,2008). A fiabilidade verifica a homogeneidade, redundância ou heterogeneidade de um instrumento, através da capacidade de reproduzir resultados, mesmo que em diferentes condições, nomeadamente, na utilização de diferentes itens para um grupo semelhante de indivíduos (consistência interna), ao longo do tempo (teste-reteste), ou entre indivíduos em diferentes ocasiões (intra-observadores) (Alexandre et al, 2013)

A fiabilidade deve ser o primeiro passo para se verificar a validação de um instrumento que, embora surja como condição necessária, não é suficiente para a mesma (Marôco & Garcia-Marques, 2006).

A análise da fiabilidade interobservador é feita para estimar possíveis erros, durante a aplicação, devido à diferença entre os observadores (teste interobservador), enquanto na fiabilidade intraobservador o mesmo observador aplica o instrumento mais de uma vez (teste-reteste). No primeiro caso, se as instruções para o uso do instrumento forem seguidas corretamente pelos dois avaliadores, os resultados devem ser consistentes entre si. No segundo caso, se o constructo a ser medido não sofrer alteração, as medidas obtidas devem ser semelhantes (Fabrício-Wehbe et al, 2013). De forma a avaliar a reprodutibilidade intra-observador de um instrumento de medida utiliza-se o *Alpha de Cronbach* (α), o coeficiente de correlação intraclasse (ICC) e o Teste de Wilcoxon; quanto à reprodutibilidade inter-observador testa-se recorrendo à analise de coeficientes de correlação e de Kappa de Cohen (Barbetta & Assis, 2008). A consistência interna, ou equivalência, é avaliada através da utilização de um teste denominado *Alpha de Cronbach* (α) e, para que a medida seja considerada adequada, os *scores* recomendados devem estar compreendidos entre o 0,7 e os 0,9 (Fitzpatrick *et al,* 1998; Silva, 2006). A fiabilidade teste-reteste pode ser calculada através do ICC e ser classificada como excelente (> 0,75), moderada (0,50 a 0,75) ou baixa (< 0,50) (Saliba *et al,* 2011).). A análise estatística com o valor de *Kappa (K)* deve estar situada entre 0,7 a 0,9 (Fitzpatrick *et al,* 1998; Silva, 2006) ou idealmente ser igual, ou superior, a 0,70 (Puga, Lopes & Costa, 2011).

A proposta de Leung, Trevena & Waters (2012) para verificar a força psicométrica do instrumento é a seguinte: os parâmetro de fiabilidade são considerados **bons** se $\alpha \geq 0.90$ para a fiabilidade interna, Coeficiente K (Landis ≥0.81 ou Fleiss > 0.75), ICC> 0.75, Correlação Pearson (r) > .95 e probabilidade de erro (p) < 0.05 na avaliação da fiabilidade intra-observado (teste/reteste)

Luís Manuel Mota de Sousa, Mestre, Enfermeiro Especialista em Reabilitação no Hospital Curry Cabral, Professor Assistente na Universidade Atlântica, Doutorando em Enfermagem na Universidade Católica Portuguesa, Investigador do CIIS, Vogal do Conselho de Enfermagem da Ordem dos Enfermeiros.

BIO

e fiabilidade inter-observador. Estes parâmetros são considerados **adequados** se α = 0.80-0.89 para a fiabilidade interna, Coeficiente K (Landis =0.61-0.80 ou Fleiss =0.60-75), ICC ICC=0.60-0.74, Correlação Pearson (r) =0.90-0.94 e probabilidade de erro (p) < 0.05 no âmbito da fiabilidade intra-observado e fiabilidade inter-observador. Os resultados são **fracos** se α = 0.70-0.79 para a fiabilidade interna, Coeficiente K (Landis =0.41-0.60 ou Fleiss =0.40-59), ICC=0.40-0.59, Correlação Pearson (r) =0.85-0.89 e probabilidade de erro (p) < 0.05 na fiabilidade intra-observado e na fiabilidade inter-observador. Os parâmetro de fiabilidade são avaliados como **muito fracos** se $\alpha \leq 0.69$ para a fiabilidade interna, Coeficiente K (Landis <0.40 ou Fleiss <0.40), ICC≤0.39, Correlação Pearson (r) ≤0.84 e probabilidade de erro (p) ≥ 0.05 na fiabilidade intra-observado e na fiabilidade inter-observador.

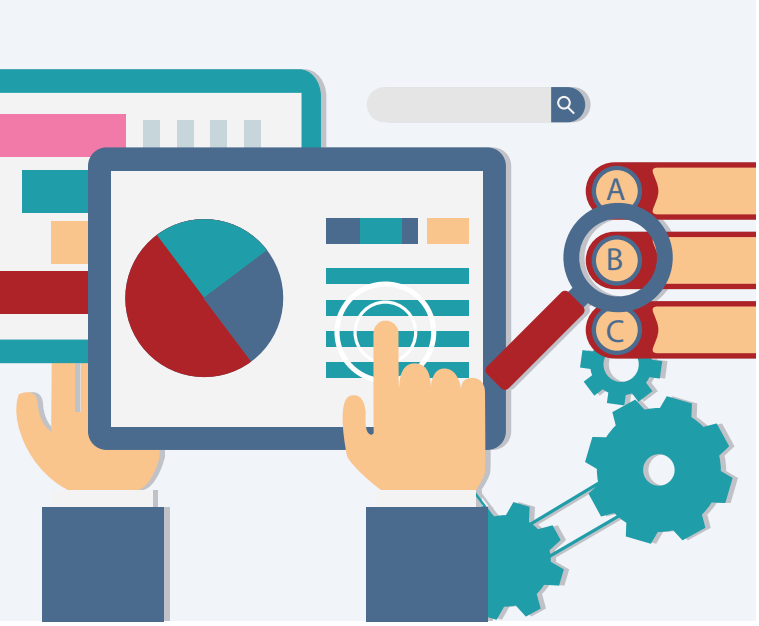
Validade

A validade permite realizar a avaliação de determinado instrumento relativamente à sua capacidade de medir aquilo a que se propõe medir (Barbetta &Assis, 2008). Esta não existe em termos absolutos a não ser no que diz respeito ao mesmo contexto/população. Existem variadas formas de estabelecer a validade de uma medida, nomeadamente: validade de conteúdo, de constructo, e de critério (Almeida et al, 2008; Fortin, 2009).

A validade de conteúdo assegura que os itens de um instrumento cobrem e representam adequadamente o que é medido (Almeida et al, 2008). Esta também pode ser definida como é a extensão em que uma medida representa todas as facetas do constructo que se pretende medir, como requisito mínimo utiliza-se um painel de peritos (Leung, Trevena & Waters, 2012). Pode ser utilizado, o coeficiente de Kappa de concordância que é a razão da proporção de vezes que os juízes concordam (corrigido pela concordância devido ao acaso), com a proporção máxima de vezes que os juízes poderiam concordar. Os valores de Kappa variam de -1 (ausência total de concordância) a 1 (concordância total) (Alexandre & Coluci, 2011). Além disso, existem outros métodos quantitativos para avaliar a força da validade de conteúdo, nomeadamente, o índice de validade de conteúdo (IVC). Este índice está relacionado com a relevância do conteúdo e classifica numa escala tipo *Likert* de quatro pontos que vai de não relevante para a altamente relevante. O limite de aceitabilidade do IVC é de 0,80, sendo considerado o mais baixo. Contudo, este índice não é amplamente divulgado e recomenda-se uma abordagem mais rigorosa para determinar a validade de conteúdo, ou seja, a relevância, a representatividade, a especificidade e a clareza dos constructos que estão a ser avaliados (Leung, Trevena & Waters, 2012).

A validade de constructo refere-se ao grau de conformidade de um instrumento com a teoria através das relações entre parâmetros importantes (Barbetta & Assis, 2008), ou seja, é uma forma quantitativa de estimar a validade de uma medida através da avaliação das relações do constructo a ser avaliado com um conjunto de outros com ele relacionado (Fortin, 2009). A validade de constructo também pode ser definida como a capacidade de o instrumento aferir um conjunto de comportamentos relacionados entre si e que se consideram estar associados ao fenómeno que está sendo medido. (Almeida et al, 2008). A validade constructo pode ser denominada como convergente/divergente ou discriminante (Leung, Trevena & Waters, 2012).

A análise factorial é a técnica mais usada para a avaliação da validade de constructo, na qual se procura verificar a validade interna do instrumento, avaliando o número de subescalas subjacentes a um conjunto de variáveis, através da análise dos componentes principais e análise multidimensional de Rasch (Ferreira & Marques, 1998). Na Análise Fatorial realizada através do método de componentes principais, ou da máxima verossimilhança, permitem selecionar as soluções que apresentem uma variância explicada total superior a 50% e, cada factor tenha pelo menos 5 a 10% da variância explicada (Marôco, 2007). A Análise de Rasch explora a dimensionalidade de um instrumento determinando se as categorias de resposta da escala podem diferenciar os participantes pelas respostas dadas, especificando a estrutura e a relação entre os indivíduos e os itens dentro de várias caraterísticas subjacentes a uma escala (Lin *et al,* 2012). Outros testes comumente utilizados são a análise de variância (Anova), as amostras de teste t (T-test), correlação Pearson (r). correlação Spearman (p), multitraço-multimétodo (Leung, Trevena & Waters, 2012).



De acordo com a proposta de Leung, Trevena & Waters (2012), os parâmetros da validade são considerados **bons** se Anova (Cohen f) ≥ 0.40; T-Test (Cohen d) ≥ 0.80 ou Eta Squared T ≥ 0.14. Correlação de Pearson e Correlação de Speraman r ou p = ±0.50-±1.0; KMO ≥0.80 percentagem total de variância ≥ 70%, P< 0.05 e α ≥0 .90. Os parâmetros da validade são considerados **adequados** se Anova (Cohen f) = 0.25-0 .39; T-Test (Cohen d) =0.50- 0.79, ou Eta Squared T = 0.06-0.13. Correlação de Pearson e Correlação de Speraman r ou p = ±0.30-±.049; KMO = 0.70-0.79 percentagem total de variância ≥ 70%, P< 0.05 e α = 0.80-0.89. Estes parâmetros são avaliados como **fracos** se Anova (Cohen f) = 0.10-0.24; T-Test (Cohen d) =0.20- 0.79, ou Eta Squared T > 0.01-0.05. Correlação de Pearson e Correlação de Speraman r ou p = ±0.10-±.29; KMO =0.60-0.69, percentagem total de variância > 70%, P< 0.05 e α = 0.70-0.79. Os parâmetros de validade são considerados **muito fracos** se Anova (Cohen f) < 0.10; T-Test (Cohen d) <0.20, ou Eta Squared T < 0.01. Correlação de Pearson e Correlação de Speraman r ou p < ±0.10; KMO =0.50-0.59, percentagem total de variância < 70%, P≥ 0.05 e α ≤ 0.69.

A validade de critério diz respeito à relação de determinada medida com outra que se considere uma medida padrão para determinado constructo. Contudo, na área da saúde e considerando medidas de estado de saúde, função ou qualidade de vida, dificilmente se encontram as medidas padrão a que a literatura se refere (Fitzpatrick *et al.,* 1998).

A validade de critério constitui-se no método mais popular para determinar validade e descreve uma relação empírica entre uma medida e um critério confiável de algum tipo. (Almeida et al, 2008). A validade critério pode ser denominada de concorrente ou preditiva (Leung, Trevena & Waters, 2012). A validade concorrente demonstra a precisão de um instrumento através da comparação com o padrão-ouro (Barbetta e Assis, 2008).

Os teste utilizados na validade critério são a análise de variância (Anova), as amostras de teste t (T-test), correlação Pearson (r). correlação Spearman (p), multitraço-multimétodo e no instrumentos de diagnóstico/risco utiliza-se a curva de ROC (área sob a curva - AUC); razão de verossimilhança positiva (LR+) e razão de verossimilhança negativa (LR-)(Leung, Trevena & Waters, 2012).

Os parâmetros da validade critério são considerados **bons** se Anova (Cohen f) ≥ 0.40, T-Test (Cohen d) ≥ 0.80 ou Eta Squared T ≥ 0.14. Correlação de Pearson e Correlação de Speraman r ou p = ±0.50-±1.0, AUC > 0.9; LR+ > 10 ou LR- < 0.10. Os parâmetros são considerados **adequados** se Anova (Cohen f) = 0.25-0 .39, T-Test (Cohen d) >0 .50- 0.79, ou Eta Squared T > 0.06-0.13. Correlação de Pearson e Correlação de Speraman r ou p = ±0.30-±.049,. AUC =0.70-0.90, LR = 5.0-10 ou LR- =0.10-0.20. Os parâmetros são considerados **fracos** se Anova (Cohen f) = 0.10-0.24, T-Test (Cohen d) = 0.20- 0.49, ou Eta Squared T = 0.01-0.05. Correlação de Pearson e Correlação de Speraman r ou p = ±0.10-±0.29, AUC =0.50-0.69, LR = 2.0-5.0 ou LR- =0.50-0.20. Por último, os parâmetros são considerados **muito fracos** se Anova (Cohen f) < 0.10, T-Test (Cohen d) < 0.20, ou Eta Squared T < 0.01, Correlação de Pearson e Correlação de Speraman r ou p < ±0.10, AUC ≤0.49, LR = 1.0-2.0 ou LR- =0.50-1.0. (Leung, Trevena & Waters, 2012).

Responsividade

A responsividade é a capacidade de um instrumento medir mudanças num período de tempo pré-estabelecido (Barbetta & Assis, 2008). A responsividade também é denominada de sensibilidade à mudança uma vez que detetar modificações, num espaço de tempo predeterminado, com o intuito de verificar a ocorrência ou não das mesmas. No entanto, não há um consenso quanto à forma de medir a sensibilidade de um instrumento embora seja uma propriedade métrica exigível cada vez mais (Fitzpatrick *et al.,* 1998).

Os indicadores de referência para avaliar a capacidade de resposta são: O *Effect Size* (ES) e o *Standardized Response Mean* (SRM). O ES determina-se pela diferença dos *scores* totais de cada avaliação, a dividir pelo desvio padrão da primeira avaliação e, o SRM apenas difere a nível do denominador e utiliza-se o desvio padrão da segunda avaliação (Fitzpatrick *et al,* 1998). O SRM corresponde à variação da mediana da primeira avaliação para a avaliação final, dividido por 3/4 da diferença da variação interquartil, entre a primeira e segunda avaliação. O Índice ES refere-se à variação da mediana da avaliação inicial para a avaliação final, dividido por 3/4 da mediana do Índice da Avaliação Inicial. Neste sentido, os maiores valores da SRM e ES indicam maior responsividade (Brasil *et al,* 2003). Além destes indicadores podem utiliza-se o efeito de teto e o efeito chão, que se definem, respetivamente, como a percentagem de indivíduos que se situam no máximo e, no mínimo do *score* possível para cada domínio (Beauséjour *et al,* 2009; Suda & Coelho, 2012). A sensibilidade de um instrumento poderá ser afetada pelo efeito de teto, e de chão, uma vez que o formato do instrumento utilizado poderá reduzir a probabilidade de melhoria, ou de agravamento, a partir de um certo ponto (Fitzpatrick *et al,* 1998).

Conclusão

Com este artigo pretendemos apresentar as propriedades métricas dos instrumentos de heteroavaliação e os seus respetivos parâmetros/indicadores, de modo a sistematizar os passos de uma validação/adequação e verificação das suas qualidades métricas.

O conhecimento sobre as propriedades psicométricas de um instrumento de heteroavaliação pode reduzir o julgamento subjetivo sobre a qualidade dos resultados que se obtêm na investigação e na prática clinica.

Recomendamos que se façam revisões sistemáticas da literatura sobre as propriedades métricas dos instrumentos e escalas que se utilizam na prática clínica, de modo a verificar se a evidência científica confirma aquele instrumento concreto como válido, fiável e responsivo.

BIBLIOGRAFIA

Alexandre, N. M. C., & Coluci, M. Z. O. (2011). Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas *Ciência & Saúde Coletiva,* 16(7), 3061-3068.

Alexandre, N. M. C., Gallasch, C. H., Lima, M. H. M., & Rodrigues, R. C. M. (2013). A confiabilidade no desenvolvimento e avaliação de instrumentos de medida na área da saúde. *Revista Eletrônica de Enfermagem,* 15(3), 800-7.

Almeida, M. H. M. D., Spínola, A. W. D. P., Iwamizu, P. S., Okura, R. I. S., Barroso, L. P., & Lima, A. C. P. D. (2008). Confiabilidade do Instrumento para Classificação de Idosos quanto à Capacidade para o Autocuidado. *Revista de Saúde Pública,* 42(2), 317-323.

Barbetta, D. D. C., & Assis, M. R. (2008). Reprodutibilidade, validade e responsividade da escala de Medida de Independência Funcional (MIF) na lesão medular: revisão da literatura. *Acta Fisiátrica,* 15(3), 176-181.

Beauséjour, M., Joncas, J., Goulet, L., Roy-Beaudry, M., Parent, S., Grimard, G., & Labelle, H. (2009). Reliability and validity of adapted French Canadian version of Scoliosis Research Society outcomes questionnaire (SRS-22) in Quebec. *Spine,* 34(6), 623-628.

Brasil, T. B., Ferriani, V.P.L., & Machado, C.S.M. (2003). Inquérito sobre a qualidade de vida relacionada à saúde em crianças e adolescentes portadores de artrites idiopáticas juvenis. *Jornal de Pediatria,* 79(1), 63-68.

Cavaco, N. S., & Alouche, S. R. (2010). Instrumentos de avaliação da função de membros superiores após acidente vascular encefálico: uma revisão sistemática. *Fisioterapia e Pesquisa,* 17(2), 178-183.

Fabrício-Wehbe, S. C. C., Cruz, I. R., Haas, V. J., Diniz, M. A., Dantas, R. A. S., & Rodrigues, R. A. P. (2013). Reprodutibilidade da versão brasileira adaptada da Edmonton Frail Scale para idosos residentes na comunidade1. *Rev. Latino-Am. Enfermagem,* 21(6), 1330-6.

Ferreira, P. L., & Marques, F. B. (1998). *Avaliação psicométrica e adaptação cultural e linguística de instrumentos de medição em saúde: princípios metodológicos gerais.* Coimbra: Centro de Estudos e Investigação em Saúde da Universidade de Coimbra.

Fitzpatrick, R., Davey, C., Buxton, M.J.,& Jones, D.R. (1998). Evaluating patient based outcome measures for use in clinical trials. *Health Technology Assessment,* 2(14):1–74.

Fortin, M. F. (2009). *Fundamentos e etapas do processo de investigação.* Loures: Lusodidacta,

Hoeman, S. P. (2000). *Enfermagem de Reabilitação: Aplicação e processo.* Loures, Lusociência. 2ª ed.

Leung, K., Trevena L. & Waters, D. (2012). Development of an appraisal tool to evaluate strength of an instrument or outcome measure. *Nurse Researcher,* 20 (2), 13-19.

Lin, K. C., Chen, H. F., Wu, C. Y., Yu, T. Y., & Ouyang, P. (2012). Multidimensional Rasch validation of the Frenchay Activities Index in stroke patients receiving rehabilitation. *Journal of Rehabilitation Medicine,* 44(1), 58-64.

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

Marôco, J., 2007. *Análise estatística com utilização do SPSS.* 3ª edição.Lisboa: Edições Sílabo.

Oliveira, A. S., & Santos, V. L. C. G. (2011). Responsividade dos instrumentos de avaliação de qualidade de vida de Ferrans e Powers: uma revisão bibliográfica. *Acta Paulista de Enfermagem,* 24(6), 839-844.

Polit, D.F., Beck C.T. & Hungle, B.P. (2011). *Fundamentos de pesquisa em enfermagem: métodos, avaliação e utilização*. 7th ed. Porto Alegre: Artmed, p. 406-26.

Puga, V. O., Lopes, A. D., & Costa, L. O. (2012). Avaliação das adaptações transculturais e propriedades de medida de questionários relacionados às disfunções do ombro em língua portuguesa: uma revisão sistemática. *Revista Brasileira de Fisioterapia,* 16, 85–93.

Saliba, V. A., Magalhães, L. D. C., Faria, C. D., Laurentino, G. E. C., Cassiano, J. G., & Teixeira-Salmela, L. F. (2011). Adaptação transcultural e análise das propriedades psicométricas da versão brasileira do instrumento Motor Activity Log. *Revista Panamericana Salud Publica,* 30(3), 262-71.

Silva, M. (2006). Medidas de resultados. *ESSfisionline,* 2 (1), 59-75.

Suda, E.Y., & Coelho, A. T. (2012). Instrumentos de avaliação para limitações funcionais associadas à instabilidade crônica de tornozelo: uma revisão sistemática da literatura. *Fisioterapia e Pesquisa,* 19(1), 79-85.