ENZO BENEVENTO
FISIOTERAPIA

ORTOPEDIA · MATERIAL DE ESTUDO

# COSMIN: Propriedades de Medida de Instrumentos Clínicos

Validade, confiabilidade, erro de medida, responsividade, mudança mínima importante e adaptação transcultural

ENZO BENEVENTO FISIOTERAPIA · NOVEMBRO DE 2026

ORTOPEDIA

# COSMIN: Propriedades de Medida de Instrumentos Clínicos

Validade, confiabilidade, erro de medida, responsividade, mudança mínima importante e adaptação transcultural

Escolher um questionário ou um teste é escolher o que se vai enxergar no paciente. Este material apresenta a taxonomia COSMIN e os critérios usados hoje para julgar cada propriedade de medida, mostra como ler erro de medida e mudança mínima importante na clínica e explica como um questionário estrangeiro é adaptado para o português.

SUMÁRIO


1. Por que avaliar propriedades de medida 3
2. A taxonomia COSMIN 4
3. Validade 5
4. Confiabilidade e erro de medida 6
5. Responsividade e mudança importante 8
6. Adaptação transcultural 10
7. Resumo e números para memorizar 11
8. Autoavaliação 12
9. Referências verificadas 13

# 1. Por que avaliar propriedades de medida

Questionários, escalas e testes clínicos são instrumentos de medida. Como qualquer instrumento, têm uma qualidade que pode ser testada: medem o que dizem medir, repetem o resultado quando nada mudou e acusam a mudança quando ela acontece. Um instrumento fraco nessas propriedades leva a decisões clínicas erradas, a uma conversa confusa com o paciente e a conclusões falsas em pesquisa.

O COSMIN (*COnsensus-based Standards for the selection of health Measurement INstruments*) é uma iniciativa internacional que, a partir de consensos entre especialistas, padronizou os nomes, as definições e os métodos de avaliação dessas propriedades. O foco original são os instrumentos relatados pelo paciente (PROMs, na sigla em inglês), como os questionários de dor e de incapacidade (Mokkink et al., 2010a). Em 2020, o grupo publicou uma ferramenta própria para estudos de confiabilidade e de erro de medida de instrumentos aplicados pelo clínico e de testes de desempenho, o que inclui medidas comuns na fisioterapia, como amplitude de movimento e testes funcionais (Mokkink et al., 2020).

### Duas avaliações diferentes

Ao ler um estudo sobre um instrumento, o COSMIN separa duas perguntas (Prinsen et al., 2018):

- **O estudo merece confiança?** O risco de viés de cada estudo é julgado com a lista de verificação COSMIN (*Risk of Bias checklist*). Cada padrão recebe muito bom, adequado, duvidoso ou inadequado, e o estudo fica com a pior nota do bloco (Mokkink et al., 2018).
- **O instrumento é bom?** O resultado do estudo é comparado com critérios de boa propriedade de medida e recebe suficiente (+), insuficiente (-) ou indeterminado (?).

Numa revisão sistemática, o conjunto dos estudos sobre cada propriedade ainda recebe um grau de qualidade da evidência (alta, moderada, baixa ou muito baixa), numa adaptação do GRADE.

> **ATENÇÃO**
>
> Até 2017, a classificação dos estudos usava os rótulos excelente, bom, razoável e pobre (Terwee et al., 2012), e eles ainda aparecem em muitas revisões. A versão atual trocou os nomes porque os antigos não correspondiam bem às descrições de cada categoria (Mokkink et al., 2018).

## 2. A taxonomia COSMIN

O consenso COSMIN organizou as propriedades de medida em três domínios; a interpretabilidade fica fora deles, porque não é propriedade de medida (Mokkink et al., 2010a).




**Figura 1.** Os três domínios da taxonomia COSMIN e a interpretabilidade, fora deles. Fonte: Mokkink et al., 2010a.

| Propriedade | Pergunta que responde |
|---|---|
| Consistência interna | Os itens de uma mesma escala andam juntos? |
| Confiabilidade | O instrumento distingue um paciente do outro, apesar do erro de medida? |
| Erro de medida | Quanto o escore varia por erro, sem mudança real? |
| Validade de conteúdo | Os itens são relevantes, completos e compreensíveis para essa população? |
| Validade estrutural | As dimensões do escore correspondem às do construto? |
| Teste de hipóteses | Os escores se relacionam com outras medidas como se espera? |
| Validade transcultural | Os itens funcionam igual em outra língua, cultura ou grupo? |
| Validade de critério | O escore concorda com um padrão-ouro? |
| Responsividade | O instrumento detecta mudança no construto ao longo do tempo? |

### A ordem de avaliação

O COSMIN recomenda começar pela validade de conteúdo, a propriedade mais importante: se os itens não são relevantes, completos e compreensíveis, o resto perde o sentido. Seguem validade estrutural, consistência interna e validade transcultural e, depois, confiabilidade, erro de medida, validade de critério, teste de hipóteses e responsividade (Prinsen et al., 2018). A viabilidade (tempo, custo, facilidade de uso) não é propriedade de medida, mas pesa na escolha entre instrumentos de qualidade parecida.

## 3. Validade

Validade é o grau em que o instrumento mede o construto que pretende medir (Mokkink et al., 2010a).

### Validade de conteúdo

Pergunta se o conteúdo do instrumento reflete adequadamente o construto. É julgada em três aspectos: relevância dos itens, abrangência (nada importante ficou de fora) e compreensibilidade (o paciente entende o item como se pretende). Exige ouvir pacientes da população-alvo, por exemplo em entrevistas cognitivas, e também profissionais (Terwee et al., 2018). A validade de face, a impressão de que o instrumento parece medir o construto, faz parte dela.

### Validade de construto

É a forma de testar a validade quando não existe padrão-ouro. Tem três aspectos (Prinsen et al., 2018):

- **Validade estrutural:** o escore reflete as dimensões do construto? É testada por análise fatorial ou por modelos de teoria de resposta ao item.
- **Teste de hipóteses:** o pesquisador declara antes que relação espera e confere. As hipóteses genéricas do COSMIN são correlação de 0,50 ou mais com instrumentos do mesmo construto, de 0,30 a 0,50 com construtos relacionados, mas diferentes, e abaixo de 0,30 com construtos sem relação. Também entram diferenças esperadas entre grupos. A validade é suficiente quando pelo menos 75% das hipóteses se confirmam.
- **Validade transcultural (invariância de medida):** os itens de uma versão traduzida, ou aplicada a outro grupo (outra idade, sexo, língua ou doença), funcionam como na versão original? É testada pela análise do funcionamento diferencial dos itens (DIF) ou por análise fatorial em vários grupos.

### Validade de critério

O escore é comparado com um padrão-ouro, no mesmo momento (validade concorrente) ou no futuro (validade preditiva). Para questionários, praticamente não existe padrão-ouro; a exceção aceita é comparar uma versão curta com a versão longa original (Mokkink et al., manual do COSMIN, 2012). Quando há padrão-ouro, o critério COSMIN é correlação de 0,70 ou mais, ou área sob a curva ROC (AUC) de 0,70 ou mais (Prinsen et al., 2018). Por isso, comparar a nota de dor com o consumo de analgésico não é validade de critério: o consumo não é padrão-ouro de dor, e a comparação entra como teste de hipóteses, com a correlação esperada declarada antes.

## 4. Confiabilidade e erro de medida

O domínio confiabilidade trata do grau em que a medida está livre de erro (Mokkink et al., 2010a). Ele reúne três propriedades: consistência interna, confiabilidade e erro de medida.

### Consistência interna

É o grau de inter-relação entre os itens de uma escala, em geral medido pelo alfa de Cronbach. Só faz sentido em escalas de modelo reflexivo, em que os itens são manifestações do mesmo construto: na ansiedade, preocupação, inquietação e pânico tendem a aparecer juntos. Numa lista de itens que não precisam andar juntos, como uma lista de comorbidades, o alfa não se aplica (Prinsen et al., 2018).

O alfa não prova que a escala mede uma coisa só. Antes dele, é preciso mostrar que a escala é unidimensional (validade estrutural); numa escala com várias dimensões, o alfa superestima a consistência (Mokkink et al., 2018). Com isso resolvido, o critério COSMIN é alfa de 0,70 ou mais em cada escala ou subescala (Prinsen et al., 2018). Um alfa baixo indica itens pouco relacionados entre si; um alfa muito alto, acima de 0,90 a 0,95, sugere itens redundantes, o que interessa sobretudo a quem constrói ou encurta um questionário (Terwee et al., 2007; Streiner, 2003).

### Confiabilidade

É a proporção da variação total dos escores que se deve a diferenças verdadeiras entre os pacientes (Mokkink et al., 2010a). É testada em pacientes estáveis, medidos duas vezes (teste-reteste), por dois avaliadores (interavaliador) ou pelo mesmo avaliador em ocasiões diferentes (intra-avaliador). Para escores contínuos usa-se o coeficiente de correlação intraclasse (ICC); para categorias ordenadas, o kappa ponderado. O critério COSMIN é ICC ou kappa ponderado de 0,70 ou mais (Prinsen et al., 2018).

Uma classificação muito usada para o ICC é a de Koo e Li. Ela deve ser aplicada ao intervalo de confiança de 95%, não só ao valor pontual, e o estudo precisa informar qual das formas de ICC usou (Koo e Li, 2016).

| ICC | Confiabilidade |
|---|---|
| Abaixo de 0,50 | Pobre |
| De 0,50 a 0,75 | Moderada |
| De 0,75 a 0,90 | Boa |
| Acima de 0,90 | Excelente |

ATENÇÃO

A confiabilidade depende da variação entre os pacientes da amostra: o mesmo instrumento tem ICC maior numa amostra heterogênea do que numa homogênea. Para acompanhar a mudança de um paciente, o que interessa é o erro de medida, na unidade do próprio instrumento (de Vet et al., 2006).

## Erro de medida

É o erro, sistemático e aleatório, do escore de um paciente que não se deve a mudança real no construto (Mokkink et al., 2010a). Ele é expresso na unidade do instrumento, de três formas:

- **Erro-padrão da medida (SEM):** o desvio-padrão do erro. Uma fórmula comum é SEM = DP × √(1 − ICC), com o desvio-padrão (DP) e o ICC da mesma amostra (Weir, 2005).
- **Menor mudança detectável (SDC, também chamada de MDC95):** SDC = 1,96 × √2 × SEM, ou cerca de 2,77 × SEM (Mokkink et al., 2010b). É a menor mudança que, num paciente, supera o erro de medida com 95% de confiança. Mudanças menores podem ser só erro.
- **Limites de concordância (LoA):** a faixa em que caem 95% das diferenças entre duas medidas do mesmo paciente estável.

EXEMPLO DE CÁLCULO

Com números ilustrativos: um questionário com DP de 10 pontos e ICC de 0,91 tem SEM = 10 × √0,09 = 3 pontos e SDC ≈ 2,77 × 3 ≈ 8,3 pontos. Uma melhora de 6 pontos ainda cabe no erro de medida; uma de 10 pontos o supera.

Para o COSMIN, o erro de medida é suficiente quando o SDC (ou os limites de concordância) é menor que a mudança mínima importante (MIC), tema da seção 5 (Prinsen et al., 2018). Se o SDC for maior, uma mudança que importa ao paciente pode ficar escondida no erro.

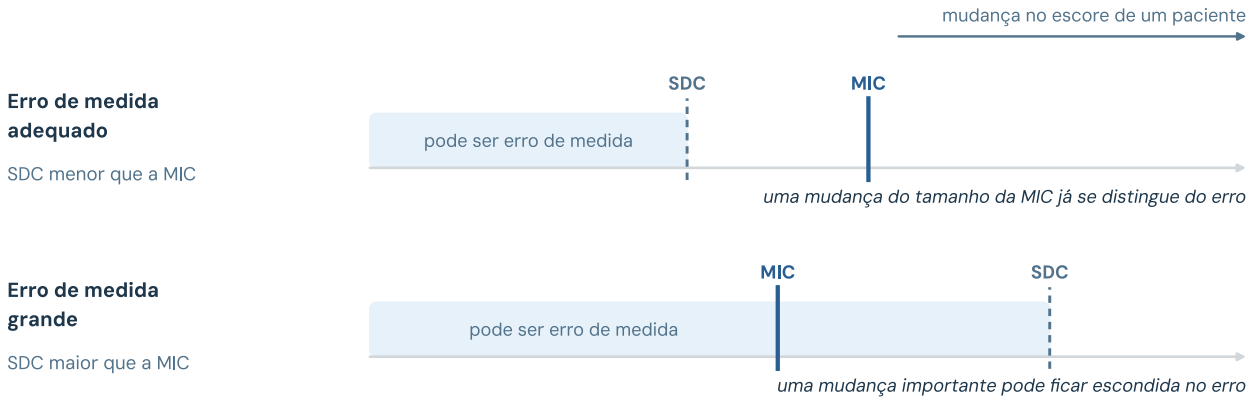



**Figura 2.** Erro de medida e mudança mínima importante no mesmo eixo: o instrumento serve para acompanhar o paciente quando o SDC é menor que a MIC.

## 5. Responsividade e mudança importante

Responsividade é a capacidade do instrumento de detectar mudança, ao longo do tempo, no construto que mede (Mokkink et al., 2010a). Para o COSMIN, é validade longitudinal: em vez de validar um escore isolado, valida-se a mudança do escore. Por isso, ela é avaliada como a validade (Angst, 2011, resposta de Mokkink et al.):

- **teste de hipóteses:** correlação esperada entre a mudança no instrumento e a mudança num instrumento de construto parecido, ou diferença esperada de mudança entre grupos;
- **comparação com um padrão-ouro de mudança,** quando ele existe: AUC de 0,70 ou mais (Prinsen et al., 2018).

A pergunta global de mudança ("Desde o início do tratamento, você está melhor, igual ou pior?") não é padrão-ouro, porque sua validade e sua confiabilidade são limitadas. Ela entra como comparador, com hipótese declarada (Mokkink et al., manual do COSMIN, 2012).

### Tamanho de efeito e SRM

O tamanho de efeito (ES) é a mudança média dividida pelo desvio-padrão inicial (Kazis et al., 1989). A média de resposta padronizada (SRM) divide a mudança média pelo desvio-padrão da mudança. Pela convenção de Cohen, 0,2 é um efeito pequeno, 0,5 moderado e 0,8 grande (Cohen, 1988).

Esses índices medem o tamanho da mudança, não a qualidade do instrumento. Um tamanho de efeito grande pode vir de um tratamento muito eficaz ou de um instrumento que exagera a mudança. Eles só contam como evidência de responsividade quando comparados com uma hipótese, declarada antes, sobre quanto efeito esperar (Terwee et al., 2003; Angst, 2011, resposta de Mokkink et al.).

DIVERGÊNCIA

Parte dos autores defende o tamanho de efeito e o SRM como medidas de responsividade, porque são fáceis de entender e sustentaram décadas de estudos (Angst, 2011). O grupo COSMIN responde que esses estudos não devem ser descartados, mas dão menos evidência do que se pensava, porque um tamanho de efeito alto não basta para dizer que o instrumento é responsivo.

### Mudança mínima importante

A mudança mínima importante (MIC) é o limiar de mudança, no mesmo paciente e ao longo do tempo, acima do qual ele se percebe importantemente mudado (Terwee et al., 2021). O conceito nasceu como diferença mínima clinicamente importante (MCID), estimada a partir de perguntas globais de mudança (Jaeschke et al., 1989).

- **Métodos baseados em âncora:** uma referência externa diz quem mudou, em geral a pergunta global de mudança. A MIC sai da mudança média no grupo que se diz “um pouco melhor”, do ponto da curva ROC que melhor separa quem melhorou de quem não mudou ou, como recomendam hoje os autores do COSMIN, de modelos preditivos (Terwee et al., 2021). Para o COSMIN, a MIC deve vir de um estudo longitudinal com âncora (Prinsen et al., 2018).
- **Métodos baseados na distribuição:** meio desvio-padrão (Norman et al., 2003), um SEM ou o SDC. Eles descrevem a variação e a precisão da medida, não o que o paciente considera importante; a menor mudança detectável não deve substituir a mudança mínima importante (de Vet e Terwee, 2010).

Na dor lombar, um consenso internacional propôs os valores abaixo para a melhora de um paciente. Levando em conta o valor inicial, uma melhora de 30% foi considerada um limiar útil em todas essas medidas. Os autores apresentam os números como ponto de partida, não como resposta final (Ostelo et al., 2008).

| Instrumento | Escala | MIC |
|---|---|---|
| Escala visual analógica (EVA) de dor | 0 a 100 mm | 15 mm |
| Escala numérica (END) de dor | 0 a 10 | 2 pontos |
| Roland-Morris | 0 a 24 | 5 pontos |
| Oswestry (ODI) | 0 a 100 | 10 pontos |
| Quebec | 0 a 100 | 20 pontos |

NA CLÍNICA

Leia a mudança de um paciente em dois degraus: acima do SDC, ela é real; acima da MIC, ela importa para o paciente. Se o SDC do instrumento for maior que a MIC, uma melhora importante pode não se distinguir do erro naquele paciente. E o número não substitui o que o paciente conta: leia também as respostas item a item.

## 6. Adaptação transcultural

Um questionário criado em outra língua precisa ser adaptado, e não apenas traduzido. O guia de Beaton et al. (2000), um dos mais usados, organiza o processo em seis etapas (Beaton et al., 2000):

1. **Tradução:** dois tradutores bilíngues, com a língua-alvo como língua materna, traduzem de forma independente; um conhece os conceitos medidos, o outro não.
2. **Síntese:** os tradutores e um observador, que registra as decisões, chegam a uma versão única.
3. **Retrotradução:** dois tradutores com a língua de origem como língua materna traduzem a versão de volta, sem acesso ao original.
4. **Comitê de especialistas:** metodologista, profissionais de saúde e de idiomas e os tradutores fecham a versão pré-final, buscando equivalência semântica (sentido das palavras), idiomática (expressões), experiencial (a atividade existe naquela cultura) e conceitual.
5. **Pré-teste:** de 30 a 40 pessoas da população-alvo respondem e contam como entenderam cada item e escolheram cada resposta.
6. **Avaliação da documentação:** os desenvolvedores ou um comitê conferem os relatórios para ver se o processo foi seguido.




**Figura 3.** As seis etapas da adaptação transcultural e o passo seguinte. Fonte: Beaton et al., 2000.

Depois do processo, o próprio guia recomenda testar as propriedades de medida da nova versão (Beaton et al., 2000). Na lista de verificação atual do COSMIN, o processo de tradução deixou de ser avaliado, porque tradução não é propriedade de medida: o pré-teste conta como evidência de validade de conteúdo (compreensibilidade), e uma tradução cuidadosa não garante validade transcultural, que se testa por DIF ou por análise fatorial em vários grupos (Mokkink et al., 2018).

PONTO-CHAVE

Uma versão em português bem traduzida ainda precisa mostrar, na população brasileira, confiabilidade, validade e responsividade. Sem esses testes, é uma tradução cuidadosa, não um instrumento validado.

## 7. Resumo e números para memorizar

| O quê | Número ou regra | Fonte |
|---|---|---|
| Qualidade do estudo | muito bom, adequado, duvidoso ou inadequado; vale a pior nota | Mokkink et al., 2018 |
| Consistência interna | alfa de 0,70 ou mais, em escala unidimensional | Prinsen et al., 2018 |
| Confiabilidade | ICC ou kappa ponderado de 0,70 ou mais | Prinsen et al., 2018 |
| Leitura do ICC | abaixo de 0,50 pobre; 0,50 a 0,75 moderada; 0,75 a 0,90 boa; acima de 0,90 excelente | Koo e Li, 2016 |
| Menor mudança detectável | SDC = 1,96 × √2 × SEM, cerca de 2,77 × SEM | Mokkink et al., 2010b |
| Erro de medida suficiente | SDC menor que a MIC | Prinsen et al., 2018 |
| Teste de hipóteses | 75% ou mais das hipóteses confirmadas | Prinsen et al., 2018 |
| Correlação com construto parecido | 0,50 ou mais | Prinsen et al., 2018 |
| Critério e responsividade com padrão-ouro | correlação ou AUC de 0,70 ou mais | Prinsen et al., 2018 |
| Tamanho de efeito | 0,2 pequeno; 0,5 moderado; 0,8 grande | Cohen, 1988 |
| MIC na dor lombar | EVA 15 mm; END 2; Roland-Morris 5; Oswestry 10; ou 30% | Ostelo et al., 2008 |
| Pré-teste da adaptação | 30 a 40 pessoas | Beaton et al., 2000 |

### Frases para fixar

- Primeiro o conteúdo: item irrelevante ou incompreensível não se salva com estatística.
- O alfa só se lê depois de mostrar que a escala é unidimensional.
- Confiabilidade distingue pacientes; erro de medida mostra quanto o escore oscila no mesmo paciente.
- Mudança menor que o SDC pode ser erro; maior que a MIC importa para o paciente.
- Tamanho de efeito mede o tamanho da mudança, não a qualidade do instrumento.
- MIC se estima com âncora; métodos de distribuição medem precisão, não importância.
- Tradução bem feita não dispensa testar as propriedades de medida na nova versão.

## 8. Autoavaliação

1. Quais são os três domínios da taxonomia COSMIN, e por que a interpretabilidade fica fora deles?
2. Que duas perguntas o COSMIN separa ao avaliar um estudo sobre um instrumento?
3. Por que a validade de conteúdo é avaliada primeiro, e quais são os seus três aspectos?
4. Quando o alfa de Cronbach pode ser interpretado, e qual é o critério COSMIN?
5. Qual é a diferença entre confiabilidade e erro de medida, e qual delas interessa para acompanhar um paciente?
6. Como se calculam o SEM e o SDC, e o que significa uma mudança menor que o SDC?
7. Por que comparar a nota de dor com o consumo de analgésico não é validade de critério?
8. Por que um tamanho de efeito grande não basta para mostrar que um instrumento é responsivo?
9. Qual é a diferença entre os métodos de MIC baseados em âncora e os baseados na distribuição?
10. Quais são as seis etapas de Beaton et al., e o que ainda falta depois delas?

**Gabarito**

1. Confiabilidade, validade e responsividade. A interpretabilidade trata de dar sentido aos escores, não da qualidade da medida.
2. Se o estudo merece confiança (risco de viés: muito bom, adequado, duvidoso ou inadequado, valendo a pior nota) e se o instrumento é bom (resultado suficiente, insuficiente ou indeterminado frente aos critérios).
3. Porque, se os itens não forem relevantes, completos e compreensíveis para o construto e a população, as outras propriedades perdem o sentido. Aspectos: relevância, abrangência e compreensibilidade.
4. Depois de mostrar que a escala ou subescala é unidimensional, e só em modelos reflexivos. Critério: alfa de 0,70 ou mais.
5. Confiabilidade é a capacidade de distinguir pacientes apesar do erro e depende da variação da amostra; erro de medida é quanto o escore varia sem mudança real, na unidade do instrumento. Para acompanhar um paciente, interessa o erro de medida.
6. SEM = DP × $\sqrt{}$(1 − ICC); SDC = 1,96 × $\sqrt{}$2 × SEM, cerca de 2,77 × SEM. Uma mudança menor que o SDC pode ser só erro de medida.
7. Porque o consumo de analgésico não é padrão-ouro de dor. A comparação serve como teste de hipóteses, com a correlação esperada declarada antes.
8. Porque ele mede o tamanho da mudança, que depende também do tratamento. Só conta como evidência quando comparado com uma hipótese, declarada antes, sobre quanto efeito esperar.
9. Os de âncora usam uma referência externa de mudança, como a pergunta global, e captam o que o paciente considera importante; os de distribuição (meio DP, um SEM, SDC) descrevem a precisão da medida e não substituem a MIC.
10. Tradução, síntese, retrotradução, comitê de especialistas, pré-teste com 30 a 40 pessoas e avaliação da documentação. Depois, falta testar as propriedades de medida da nova versão.

## 9. Referências verificadas

Consultadas em 19 de novembro de 2026.


1. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-745. doi:10.1016/j.jclinepi.2010.02.006
2. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539-549. doi:10.1007/s11136-010-9606-8
3. Mokkink LB, Terwee CB, Patrick DL, et al. COSMIN checklist manual. Amsterdam: VU University Medical Center; 2012.
4. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012;21(4):651-657. doi:10.1007/s11136-011-9960-1
5. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018;27(5):1171-1179. doi:10.1007/s11136-017-1765-4
6. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147-1157. doi:10.1007/s11136-018-1798-3
7. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159-1170. doi:10.1007/s11136-018-1829-0
8. Mokkink LB, Boers M, van der Vleuten CPM, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med Res Methodol. 2020;20(1):293. doi:10.1186/s12874-020-01179-5
9. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42. doi:10.1016/j.jclinepi.2006.03.012
10. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003;80(1):99-103. doi:10.1207/S15327752JPA8001_18
11. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
12. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033-1039. doi:10.1016/j.jclinepi.2005.10.015
13. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231-240. doi:10.1519/15184.1

14. Terwee CB, Dekker FW, Wiersinga WM, et al. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res. 2003;12(4):349-362. doi:10.1023/a:1023499322593
15. Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness (com resposta de Mokkink LB, Terwee CB, Knol DL, de Vet HCW). BMC Med Res Methodol. 2011;11:152. doi:10.1186/1471-2288-11-152
16. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989;27(3 Suppl):S178-S189. doi:10.1097/00005650-198903001-00015
17. Cohen J. Statistical power analysis for the behavioral sciences. 2. ed. Hillsdale: Lawrence Erlbaum; 1988.
18. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10(4):407-415. doi:10.1016/0197-2456(89)90005-6
19. Terwee CB, Peipert JD, Chapman R, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res. 2021;30(10):2729-2754. doi:10.1007/s11136-021-02925-y
20. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003;41(5):582-592. doi:10.1097/01.MLR.0000062554.74615.4C
21. de Vet HCW, Terwee CB. The minimal detectable change should not replace the minimal important difference. J Clin Epidemiol. 2010;63(7):804-805. doi:10.1016/j.jclinepi.2009.12.015
22. Ostelo RW, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine. 2008;33(1):90-94. doi:10.1097/BRS.0b013e31815e3a10
23. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine. 2000;25(24):3186-3191. doi:10.1097/00007632-200012150-00014



ENZO BENEVENTO

FISIOTERAPIA