UNIVERSIDADE FEDERAL DO AMAZONAS

PROGRAMA DE PÓS-GRADUAÇÃO STRICTO SENSU EM CIÊNCIAS DO MOVIMENTO HUMANO

WALBERT MENEZES BITAR

PROPRIEDADES DE MEDIÇÃO DOS TESTES DE DESEMPENHO FÍSICO FUNCIONAL USADOS EM PESSOAS IDOSAS: UMA REVISÃO SISTEMÁTICA SEGUINDO AS DIRETRIZES COSMIN

MANAUS
2023

WALBERT MENEZES BITAR

PROPRIEDADES DE MEDIÇÃO DOS TESTES DE DESEMPENHO FÍSICO FUNCIONAL USADOS EM PESSOAS IDOSAS: UMA REVISÃO SISTEMÁTICA SEGUINDO AS DIRETRIZES COSMIN

Dissertação apresentada como trabalho de conclusão de Mestrado para o Programa de Pós-Graduação Stricto Sensu em Ciências do Movimento Humano (PPGCiMH) da Faculdade de Educação Física e Fisioterapia da Universidade Federal do Amazonas (FEFF/UFAM) na linha de pesquisa avaliação e recuperação funcional, como requisito parcial à obtenção do título de mestre em Ciências do Movimento Humano.

Orientador: Prof.º Dr Ewertton de Souza Bezerra

MANAUS
2023

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Bitar, Walbert Menezes
B624p Propriedades de medição dos testes de desempenho físico funcional usados em pessoas idosas : uma revisão sistématica seguindo as diretrizes COSMIN / Walbert Menezes Bitar . 2023
106 f.: il. color; 31 cm.

Orientador: Ewertton de Souza Bezerra
Dissertação (Mestrado em Ciências do Movimento Humano) - Universidade Federal do Amazonas.

1. Envelhecimento. 2. Propriedades de medição. 3. Validade. 4. Confiabilidade. 5. Desempenho físico e funcional. I. Bezerra, Ewertton de Souza. II. Universidade Federal do Amazonas III. Título

WALBERT MENEZES BITAR

PROPRIEDADES DE MEDIÇÃO DOS TESTES DE DESEMPENHO FÍSICO FUNCIONAL USADOS EM PESSOAS IDOSAS: UMA REVISÃO SISTEMÁTICA SEGUINDO AS DIRETRIZES COSMIN

Aprovado em

Banca examinadora

_____

Prof. Dr. Ewertton de Souza Bezerra (PPGCiMH-FEFF-UFAM)

Presidente

_____

Prof. Dr. João Otacílio Libardoni dos Santos (PPGCiMH-FEFF-UFAM)

Membro Interno

_____

Prof. Dr. Walan Robert da Silva (UNIAVAN)

Membro Externo

MANAUS
2023

Dedicatória

Dedico este trabalho a todos que me ajudaram nessa caminhada superando obstáculos. Em especial, à minha esposa Mary Jully, meus filhos Wandrew e Wayola. À minha mãe Waldenôra e ao meu pai Romildo (em memória).

AGRADECIMENTOS

À Universidade Federal do Amazonas, juntamente ao Programa de Pós-Graduação de Ciências do Movimento Humano (PPGCiMH) por fazer parte da primeira turma de mestrado da Faculdade de Educação Física e Fisioterapia. À Fundação de Amparo e Pesquisa do Amazonas (FAPEAM), e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo fomento ao PPGCiMH.

Aos professores do curso por todo o conhecimento transmitido de forma cientifica e responsável, em especial, ao meu orientador Prof. Dr. Ewertton Bezerra, que pelos seus esforços acadêmico e profissional, se tornou uma referência, ansiando-me a seguir seus passos desde as aulas de pós-graduação lato sensu e cursos de extensão até a conquista do ingresso no mestrado, onde firmou, acima de tudo no respeito, uma parceria altruísta até agora, meu muito obrigado.

Ao grande amigo que a profissão me deu Prof. Me. Aluísio Avelino, aos amigos de mestrado Jean Carlos e Geovanna Souza, agradeço demais pela ajuda nas etapas iniciais deste trabalho, no entanto, gostaria de agradecer em especial a querida Profa. Me. Suzy Pinto, com certeza contribuiu em maior magnitude e foi peça fundamental na realização do trabalho, auxiliando-me nas etapas finais e, mesmo residindo em outro país, se fez tão presente que a distância foi superada por sua presteza, muito obrigado meus amigos pela solidariedade.

Aos familiares, pois mais do que a motivação extrínseca para seguir com os objetivos, foram e são minha base para que eu pudesse continuar nas inúmeras vezes que os obstáculos apareceram ao longo dessa jornada, muito obrigado a todos vocês sem exceção.

Mas agradeço especialmente a quem esteve ligado diretamente nesse processo que, muitas vezes, nos faz ficar distantes dos mais próximos e sem a compreensão dos mesmos, é impossível alcançar o objetivo final, portanto, meu muito obrigado a minha esposa Mary Jully e meus filhos Wandrew e Wayola, foi, é e sempre será por vocês.

Finalmente, mas não menos importante, agradeço a Deus pai pela vida de todas as pessoas citadas neste agradecimento, sem as quais estes agradecimentos não teriam existido porque contribuíram de alguma forma para a realização deste sonho, devo tudo a ti senhor.

“Ao expandirmos o campo do conhecimento, apenas aumentamos o horizonte da ignorância”

Henry Miller

RESUMO

Introdução: Para as pessoas idosas, a avaliação do desempenho físico funcional tem sido um dos pilares da avaliação clínica, pois capacidade funcional norteada pela capacidade intrínseca e pelo meio ambiente, é pedra angular para ao envelhecimento bem-sucedido. A capacidade intrínseca afetada pelo envelhecimento inclui funções sensoriais, cognitivas e de movimento, as medidas de capacidade intrínseca dentro de cada uma dessas funções são avaliadas por meio de testes de desempenho físico funcional. No entanto, não há uma análise criteriosa sobre as propriedades de medição destes testes, tais como: a validade e a confiabilidade. **Objetivo:** Avaliar as propriedades de medição dos testes que mensuram o desempenho físico funcional de pessoas idosas com o foco nos membros inferiores, isto é: *Timed Up and Go Test* (*TUG*), Teste de Sentar e Levantar da Cadeira (TSLC), Teste de Velocidade de Caminhada (TVC) e Teste de Potência para Membros Inferiores (TPMI) por meio de uma visão sistemática. **Métodos:** As bases de dados *PubMed*, *Embase* e Biblioteca Virtual em Saúde (BVS) foram sistematicamente pesquisadas, para localizar estudos sobre as propriedades de medição (validade e confiabilidade), dos testes de desempenho físico funcional usados em pessoas idosas da comunidade e institucionalizados com idade acima de 60 anos. O manual do usuário *COSMIN methodology for systematic reviews* foi usado para avaliação da qualidade dos estudos incluídos. **Resultados:** Foram incluídos 18 estudos, com oito estudos correspondendo ao teste TUG, três estudos o TSLC, nove estudos o TVC e quatro estudos o TPMI. Em 15 estudos foi avaliado o domínio confiabilidade, 13 estudos avaliaram erro de medição, dois estudos avaliaram o teste de hipóteses para validade de construto e oito estudos avaliaram a validade de critério. A avaliação da qualidade geral das propriedades de medição e das evidências foram bastante variáveis e, em geral, abaixo do ideal com melhor resultados para confiabilidade entre os testes. **Conclusão:** Há uma clara necessidade de maior atenção a qualidade nos processos de construção, adaptação e validação das propriedades de medição dos testes de desempenho físico funcional usados em pessoas idosas. No entanto, a certeza das evidências foi avaliada como moderadas e baixas.

Palavras-chave: Envelhecimento, Propriedades de medição, Validade, Confiabilidade, Avaliação, Desempenho físico e funcional.

ABSTRACT

Introduction: For the elderly, the assessment of functional physical performance has been a cornerstone of clinical evaluation, as functional capacity, guided by intrinsic capacity and the environment, is a cornerstone for successful aging. Intrinsic capacity affected by aging includes sensory, cognitive, and movement functions, and measures of intrinsic capacity within each of these functions are assessed through functional physical performance tests. However, there is a lack of thorough analysis of the measurement properties of these tests, such as validity and reliability. **Objective:** Evaluate the measurement properties of tests that assess the functional physical performance of the elderly, focusing on the lower limbs, namely: the Timed Up and Go Test (TUG), Chair Sit-to-Stand Test (CSST), Walking Speed Test (WST), and Lower Limb Power Test (LLPT), through a systematic view. **Methods:** PubMed, Embase, and the Virtual Health Library (VHL) databases were systematically searched to locate studies on the measurement properties (validity and reliability) of functional physical performance tests used in community-dwelling and institutionalized elderly people aged 60 and over. The COSMIN methodology for systematic reviews user manual was used to assess the quality of the included studies. **Results:** Eighteen studies were included, with eight studies corresponding to the TUG test, three studies to the CSST, nine studies to the WST, and four studies to the LLPT. Fifteen studies evaluated the reliability domain, 13 studies assessed measurement error, two studies assessed the construct validity hypothesis test, and eight studies evaluated criterion validity. The overall assessment of the quality of the measurement properties and evidence was quite variable and generally below ideal, with better results for reliability between tests. **Conclusion:** There is a clear need for greater attention to quality in the construction, adaptation, and validation processes of the measurement properties of functional physical performance tests used in the elderly. However, the certainty of evidence was assessed as moderate to low.

Keywords: Aging, Measurement Properties, Validity, Reliability, Assessment, Physical and Functional Performance.

APRESENTAÇÃO

Esta dissertação de conclusão do mestrado está dividida em 14 tópicos: introdução, revisão da literatura, materiais e métodos, resultados, discussão, conclusão, referências e os apêndices além dos elementos complementares: resumo, listas de abreviações, siglas, quadros, figuras e tabelas. Para facilitar o entendimento do leitor, cada um destes tópicos tem a seguinte ordem:

1. Introdução de forma discorrida, há uma contextualização do escopo temático, principais lacunas, justificativas e problemática além dos objetivos, que estão subdivididos com o objetivo geral e os objetivos específicos estando atrelados ao problema de pesquisa.

2. Uma breve revisão da literatura sobre a origem e características dos testes de desempenho físico funcional (*TUG*, TSLC, TVC e TPMI).

3. Os materiais e métodos, estão todos os procedimentos detalhados que foram realizados na pesquisa (estratégia de busca, critérios de elegibilidade, síntese e extração dos dados, avaliação da qualidade metodológica e avaliação da qualidade geral da evidência.

4. Finalmente os resultados da pesquisa, onde são apresentados o fluxograma do processo de busca e seleção dos estudos e as tabelas de todo o processo descrito no tópico dos materiais e métodos.

5. Na discussão, é apresentado uma interpretação geral dos resultados, pontos fortes e limitações da revisão.

6. Por fim, na conclusão é salientado as implicações dos resultados para a prática e pesquisas futuras.

7. Neste tópico constam as 94 referências que deram todo aporte teórico para construção desta revisão.

8. O apêndice A, é apresentado a estratégia de busca realizada na base de dados *PubMed*, que seguiu para as demais bases de dados escolhidas com os ajustes necessários para cada uma.

9. O apêndice B, é apresentado o *checklist PRISMA* (2020), que apesar de ser fortemente recomendado como um norteador de redação em revisões sistemáticas, há alguns itens que não são correspondentes ao tipo da nossa revisão, por isso, alguns itens não foram reportados e os itens 24 a 27 na sessão “outra informação” serão reportados no modelo de artigo.

10 a 13. Os apêndices C, D, E e F são referentes as etapas de avaliações cegas pelos revisores, disponibilizados como materiais suplementares que devem ser acessados por links.

14. O Apêndice G exibe o link e a capa do livro digital "Avaliação Físico Funcional em Idosos", no qual tive a oportunidade de contribuir por meio deste trabalho.

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

p. ex.	por exemplo
vs.	versus
s	segundos
cm	centímetro
m	metro
m/s	metros por segundos
n/a	não aplicável
kg	quilograma
W	watts
Hz	hertz
GA	GymAware
DF	Dartfish

LISTA DE SIGLAS

OMS	Organização Mundial da Saúde
TUG	Timed Up and Go Test
TSLC	Teste de Sentar e Levantar da Cadeira
TVC	Teste de Velocidade de Caminhada
TPMI	Teste de Potência de Membros inferiores
ABVD	Atividades Básicas da Vida Diária
AIDV	Atividades Instrumentais da Vida Diária
AVD	Atividades da Vida Diária
SPPB	Short Physical Performance Battery
PADL	Performance Test of Activities of Daily Living
UEPPB	Upper Extremity Physical Performance Battery
LEPPB	Lower Extremity Physical Performance Battery
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
PROSPERO	International Prospective Register of Systematic Reviews
PROMs	Patient-Reported Outcome Measures
BVS	Biblioteca Virtual em Saúde
PUBMED	Search Engine With Free Acess to the Database
EMBASE	Comprehensive Medical Research Database
MESH	Medical Subject Headings
ICC	Intraclass Correlation Coefficient
CI	Confidence Interval
SEM	Standard Error of Measurement
CV	Coefficient of Variation
MIC	Minimal Important Change,
SDC	Smallest Detectable Change
MDC	Minimal Detectable Change
SRD	Smallest Real Difference
MDD	Minimal Detectable Difference
LoA	Limits of Agrément
AUC	Area Under the Curve

TEM	Typical Measurement Error
STSp	Sit-to-Stand Power Test
STS	Sit-to-Stand Test
5xSTS	5 Times Sit-to-Stand Test
FTST	Five Times Sit-to-Stand Test
LP	Leg Press
CMJ	Counter Movement Jump
TGUG	Timed Get Up and Go Test
Cr	Five Chair Rises
MVC	Maximum Voluntary Isometric Contraction
NGS	Normal Gait Speed
FGS	Fast Gait Speed
WS	Walking Speed
MWS	Maximum Walking Speed
SSWS	Self-Selected Walking Speed
TVC4M	Teste de Velocidade de Caminhada de 4 Metros
TVC10M	Teste de Velocidade de Caminhada de 10 Metros
3MBWT	Three-Meter Back Ward Walk Test
50FWT	50-ft Walk Test
LPT	Linear Position Transducer
IMU	Inertial Measurement Unit

SUMÁRIO

1 INTRODUÇÃO

Em 2015, a Organização Mundial da Saúde (OMS) divulgou seu primeiro relatório sobre aspectos da saúde no envelhecimento, e reconheceu à verdadeira relevância da capacidade funcional definida como "o conjunto de todas as capacidades físicas e mentais que um indivíduo pode utilizar somado ao meio que ele vive resumida em dois fatores, capacidade intrínseca e meio ambiente". Definindo o envelhecimento saudável e bem-sucedido como o desenvolvimento e a manutenção da capacidade funcional, ou seja, o estado de saúde do idoso é definido pelo estado funcional e não pela morbidade (WORLD HEALTH ORGANIZATION, 2015, p. 13).

A capacidade intrínseca afetada pelo envelhecimento inclui funções sensoriais, cognitivas e de movimento, as medidas de capacidade intrínseca dentro de cada uma dessas funções são avaliadas por meio de testes chamados de testes de capacidades, equivalentes aos testes de desempenho físico funcional, são padronizados, com medidas objetivas usando critérios predeterminados, que podem incluir cronometragem, contagem de repetições de movimentos ou tarefas como locomoção com mudança de direção (GURALNIK *et al.*, 1989; LAMB; KEENE, 2017; WORLD HEALTH ORGANIZATION, 2015).

Nesse contexto, o rastreio e a avaliação para detectar precocemente o declínio funcional ou incapacidade é um fator-chave, pois ao identificar as pessoas idosas de acordo com o estágio do desempenho físico funcional, provavelmente facilitaria a personalização dos programas de intervenção para otimizar a meta do envelhecimento bem-sucedido (ANGULO *et al.*, 2020). Por isso, as medidas de desempenho físico funcional são de suma importância, elas fornecem estimativas objetivas do aparelho locomotor no aspecto físico, auxiliando na orientação para o tratamento geriátrico e/ou avaliação da eficácia de tratamentos (GILL, 2010; GURALNIK *et al.*, 2000).

Essa importância se deve principalmente à grande evidência de que as medidas de desempenho físico funcional são mais rápidas, portáteis e menos influenciadas por origens culturais e educacionais, em comparação com medidas de autorrelato (BERNABEU-MORA *et al.*, 2015; GOBBENS; VAN ASSEN; SCHALK, 2014). Do ponto de vista prático, a viabilidade, os equipamentos necessários e os aspectos econômicos representam fatores importantes para a escolha ou não de um teste. De uma perspectiva científica, no entanto, os testes devem possuir níveis apropriados em suas propriedades de medição a fim de serem usados com confiança e serem capazes de tirar conclusões significativas dos resultados dos testes (CURRELL; JEUKENDRUP, 2008; ROBERTSON *et al.*, 2017; TERWEE *et al.*, 2007).

Explorar as propriedades de medição, como validade e confiabilidade, é um passo crucial para escolhas ou até mesmo o desenvolvimento futuro mais confiáveis de um conjunto de ferramentas para rastrear ou diagnosticar o declínio da função física e funcional e consequentemente identificar precocemente pessoas idosas com risco aumentado de resultados adversos, como incapacidade física, maior propensão a quedas, perda de independência e mortalidade.

Duas revisões publicadas existentes concentram-se apenas em pessoas idosas da comunidade. A primeira revisão é a de Freiberger *et al.*, (2012), restringiu as buscas apenas à língua inglesa e focou somente em sumarizar estudos que avaliaram o desempenho físico funcional por meio de bateria de testes que adotam um escore ou índice geral de várias dimensões físicas, como força, potência, equilíbrio, agilidade, coordenação, o que implica em um tempo de realização maior e consequentemente na praticidade dos testes. Ademais, a variedade de pontuações compostas por meio de bateria de testes que usam essas medidas em combinação, não deixa claro se os resultados com essas pontuações são impulsionados por uma medida em particular ou se cada uma delas têm contribuições aditivas semelhantes (COOPER *et al.*, 2010).

A segunda revisão de Mijnarends *et al.*, (2013), limitou as buscas pelos idiomas inglês e holandês resultando em estudos que utilizaram instrumentos de medida para medir massa muscular, força e desempenho físico. Além disso, foi avaliada a viabilidade de tais ferramentas somente em pessoas idosas da comunidade e a identificação de conjuntos de ferramentas mais válidas e confiáveis para apoiar o desenvolvimento de uma ferramenta de triagem para sarcopenia nesta mesma população. Ademais, a presente revisão compilou estudos que avaliaram o desempenho físico funcional por meio de baterias de testes, medidas de autorrelatos, medidas por meio de vídeo e medidas objetivas que abordam uma ou mais dimensões físicas, variando os tipos de instrumentos de medida.

Sendo assim, resta analisar as propriedades de medição de testes que utilizam somente medidas objetivas mais práticos e acessíveis com o mínimo de tarefas possíveis, mesmo que abordem apenas uma dimensão física com o foco principalmente no membro inferior e, como os resultados se aplicam as pessoas idosas institucionalizadas, uma vez que, pessoas idosas que residem em instituições de longa permanência, por exemplo, estão negativamente correlacionado com o nível de atividade física, multimorbidades e com alto risco de início ou progressão da incapacidade, sugerindo que pessoas idosas que vivem na comunidade teriam melhor desempenho nas avaliações de desempenho físico funcional (CHAD *et al.*, 2005; CSAPO; GORMASZ; BARON, 2009; FERRUCCI *et al.*, 2004; LEÓN-SALAS *et al.*, 2015).

1.1 Problema de pesquisa

As propriedades de medição dos testes de desempenho físico funcional de medidas objetivas são apropriadas a ponto de não comprometer a análise do desempenho nos testes em pessoas idosas da comunidade e institucionalizadas?

1.2 Justificativa

É fato bastante conhecido que os fenômenos populacionais (redução das taxas de fecundidade, mortalidade e aumento da expectativa de vida) estão impactando sobre a estrutura etária no mundo. O desequilíbrio desta balança etária produz consequências de ordem demográfica, socioeconômica, e uma das mais importantes diz respeito às demandas de saúde, ocasionada pelo crescimento do número de pessoas idosas resultando em um novo perfil epidemiológico de atenção à saúde. Pois o envelhecimento é um processo biológico que sozinho gera o declínio das capacidades físicas podendo causar doenças e síndromes geriátricas, capacidade funcional ruim e quando associado ao mal que assola o mundo contemporâneo "sedentarismo" torna este processo ainda mais severo.

Diante deste cenário, é imprescindível a utilização dos testes de desempenho físico funcional na avaliação rotineira da capacidade funcional de pessoas idosas, esta, é definida como a capacidade das pessoas realizarem seus atributos ligados a saúde afim de assegurar seu espaço social, ou seja, é a combinação de indivíduos e seus ambientes e a interação entre eles. Outros autores conceituam como um termo usado para descrever um indivíduo com a capacidade independente de realização das ABVD e/ou AIVD, sendo este um resultado significativo e profundo para a saúde de pessoas idosas.

No entanto, escolher um teste não é uma tarefa fácil, pois há inúmeros testes propostas na literatura com inúmeras informações que mais geram dúvidas do que esclarecimentos, tais como: qual o teste mais apropriado para avaliar cada domínio físico; o teste se aplica para pessoas idosas da comunidade e institucionalizadas; o teste se aplica para qualquer faixa de idade; qual a correta execução do teste; como os testes foram validados e se eles são confiáveis etc. Logo, a presente revisão busca esclarecer dúvidas como as supracitadas e outras possíveis para facilitar a escolha dos testes mais práticos, acessíveis e confiáveis.

Pois o declínio da capacidade física com o envelhecimento pode ser amplamente revertido por exercícios adequados e intervenção nutricional, e ambos podem ter suas prescrições mais assertivas quando norteados por avaliações de rastreio mais precisas e confiáveis promovendo não só um envelhecimento mais saudável como uma possível diminuição com tratamentos medicamentosos e gastos com a atenção médico-hospitalar.

1.3 Objetivos

1.3.1 Objetivo geral

- Avaliar as propriedades de medição dos testes de desempenho físico funcional usados em pessoas idosas com o foco nos membros inferiores por meio das diretrizes COSMIN para revisões sistemáticas.

1.3.2 Objetivos específicos

- Identificar e comparar as propriedades de medição dos testes de desempenho físico funcional aplicados em pessoas idosas;
- Avaliar o domínio confiabilidade dos testes de desempenho físico funcional, isto é, propriedade de medição teste-reteste, propriedade de medição intra e entre avaliadores, dias ou sessões, ensaios e o erro de medição;
- Avaliar o domínio validade, ou seja, propriedade de medição validade de construto por meio do teste de hipóteses como aspecto desta propriedade, além da propriedade de medição validade de critério;
- Avaliar a qualidade metodológica dos estudos incluídos para o *TUG* utilizando o *COSMIN Risk of Bias Checklist*;
- Avaliar a qualidade metodológica dos estudos incluídos para o *TSLC* utilizando o *COSMIN Risk of Bias Checklist*;
- Avaliar a qualidade metodológica dos estudos incluídos para o *TVC* utilizando o *COSMIN Risk of Bias Checklist*;
- Avaliar a qualidade metodológica dos estudos incluídos para o *TPMI* utilizando o *COSMIN Risk of Bias Checklist*;
- Avaliar as propriedades de medição dos estudos incluídos para o *TUG* de acordo com os critérios para as boas propriedades de medição estabelecidos pelo *COSMIN methodology for systematic reviews*;
- Avaliar as propriedades de medição dos estudos incluídos para o *TSLC* de acordo com os critérios para as boas propriedades de medição estabelecidos pelo *COSMIN methodology for systematic reviews*;
- Avaliar as propriedades de medição dos estudos incluídos para o *TVC* de acordo com os critérios para as boas propriedades de medição estabelecidos pelo *COSMIN methodology for systematic reviews*;

- Avaliar as propriedades de medição dos estudos incluídos para o *TPMI* de acordo com os critérios para as boas propriedades de medição estabelecidos pelo *COSMIN methodology for systematic reviews*;
- Avaliar as propriedades de medição sumarizadas do teste *TUG* de acordo com o *COSMIN methodology for systematic reviews*;
- Avaliar as propriedades de medição sumarizadas do teste *TSLC* de acordo com o *COSMIN methodology for systematic reviews*;
- Avaliar as propriedades de medição sumarizadas do teste *TVC* de acordo com o *COSMIN methodology for systematic reviews*;
- Avaliar as propriedades de medição sumarizadas do teste *TPMI* de acordo com o *COSMIN methodology for systematic reviews*;
- Avaliar a qualidade geral da evidência para as propriedades de medição do *TUG* pela abordagem modificada *GRADE*;
- Avaliar a qualidade geral da evidência para as propriedades de medição do *TSLC* pela abordagem modificada *GRADE*;
- Avaliar a qualidade geral da evidência para as propriedades de medição do *TVC* pela abordagem modificada *GRADE*;
- Avaliar a qualidade geral da evidência para as propriedades de medição do *TPMI* pela abordagem modificada *GRADE*;

2 REVISÃO DA LITERATURA

Para as pessoas idosas, a avaliação da função física e da incapacidade tem sido um dos pilares da avaliação clínica e existem razões convincentes para que isso aconteça. Como já mencionado, a qualidade de vida das pessoas idosas é julgada mais pelo seu nível funcional e capacidade de manter-se independente do que pelas doenças específicas diagnosticadas pelo seu médico (GURALNIK *et al.*, 1989; KATZ *et al.*, 1963; WORLD HEALTH ORGANIZATION, 2015). Adicionalmente, duas importantes limitações físicas são o funcionamento neuromotor (p. ex. resistência, força e flexibilidade articular) e o funcionamento sensorial (p. ex. visão e audição). Para a avaliação desses domínios, duas abordagens podem ser utilizadas: medidas subjetivas de autorrelato e testes objetivos baseados em desempenho (BRANCH; MEYERS, 1987; KEMPEN *et al.*, 1996; MYERS *et al.*, 1993).

O uso de escalas e índices (Índice de Barthel, Índice de Katz, Índice de Pfeffer, Índice de Lawton-Brody) criados e aprimorados com o intuito de avaliar o funcionamento autorrelatado tem uma longa história e ampla aplicação. As escalas de avaliação foram desenvolvidas a partir das ABVD e tarefas mais complexas como AIVD (FEINSTEIN; JOSEPHY; WELLS, 1986; LAWTON; BRODY, 1969). Medidas de ABVD foram introduzidas pela primeira vez como um meio sistemático de avaliar pacientes com doenças crônicas por meio de uma escala graduada e denominada "Índice de Independência nas Atividades da Vida Diária" (Índice AVD) (KATZ *et al.*, 1963).

O Índice de AVD foi desenvolvido a partir da observação de muitas atividades realizadas por um grupo de pacientes com fratura de quadril, permitindo classificar os indivíduos de acordo com adequação do desempenho. A adequação é expressa como uma nota (A, B, C, D, F, G ou outro) que resume o desempenho geral em seis tarefas: tomar banho, vestir-se, ir ao banheiro, transferir-se, continência e alimentar-se (MULTIDISCIPLINARY STUDIES OF ILLNESS IN AGED PERSONS. II. A NEW CLASSIFICATION OF FUNCTIONAL STATUS IN ACTIVITIES OF DAILY LIVING, 1959, p. 55 a 57).

A aplicação dessas medidas de funcionamento, para a época, a qual não havia medidas válidas e bem definidas, provou ser uma ferramenta valiosa tanto na pesquisa do envelhecimento quanto no atendimento clínico de pessoas idosas. No entanto, uma série de limitações metodológicas importantes foram descritas para instrumentos que avaliam o funcionamento físico por meio de autorrelato ou proxy (FEINSTEIN; JOSEPHY; WELLS, 1986; SCHULER; MARZILLI, 2003; SUCHY; KRAYBILL; FRANCHOW, 2011).

Como não existe um padrão-ouro, é difícil avaliar a validade direta das medidas de funcionamento, pois a avaliação do funcionamento ou estado geral de saúde de um indivíduo pode variar consideravelmente a partir da avaliação de um familiar ou profissional de saúde. Por exemplo, quando se pergunta aos indivíduos se eles precisam de ajuda para comer ou tomar banho, pode não estar claro quais tarefas específicas dentro do conjunto desses comportamentos.

Além disso, quando não são fornecidas diretrizes para avaliar o nível de dificuldade, pode ser problemático para os sujeitos relatarem se têm pouca, alguma ou muita dificuldade para uma atividade como caminhar uma distância especificada, dando margem para um falso relato da parte do idoso, de um membro da família, de seu cuidador ou até mesmo de um profissional da saúde com descompassos nas taxas de concordâncias de superestimação ou subestimação do estado funcional do idoso em algumas tarefas (FIGUEREDO; JACOB-FILHO, 2018; MAGAZINER *et al.*, 1997).

Por conseguinte, o declínio pode precisar atingir uma certa magnitude para poder interferir no funcionamento diário antes que possa ser reconhecido como um problema. Por esta razão, medidas de autorrelato não se mostraram sensíveis para detectar mudanças de pequena magnitude chamadas de "deficiência pré-clínica" (BRACH *et al.*, 2002).

Nesse contexto, estudos longitudinais e algumas revisões mostraram mudanças substanciais no funcionamento autorrelatado e por *proxy* ao longo do tempo, o que pode ser devido a uma mudança real, mas provavelmente também resulta de algum grau de falta de confiabilidade nas medidas, ou seja, a medição de ABVD, apenas no formato de pergunta, geralmente tem baixa confiabilidade, validade, reprodutibilidade e sensibilidade à mudança (BRACH *et al.*, 2002; DE MORTON; BERLOWITZ; KEATING, 2008; FIEO *et al.*, 2011; HOPMAN-ROCK *et al.*, 2019). Além do mais, muitos estudos comparando diferentes medidas de avaliação do estado funcional do idoso (autorrelato e medidas baseadas em desempenho) normalmente encontraram uma baixa ou razoável, mas não grande correlação (REUBEN *et al.*, 1995; SCHULER; MARZILLI, 2003; SHERMAN; REUBEN, 1998; SUCHY *et al.*, 2010).

Em contrapartida, medidas objetivas por meio de avaliação do desempenho físico, oferecem o potencial de maior reprodutibilidade. Em um estudo com indivíduos com problemas respiratórios e cardíacos, o teste de caminhada de 6 minutos, repetido seis vezes em intervalos de duas semanas, mostrou-se mais reprodutível do que quatro diferentes questionários de estado funcional e dois testes de função pulmonar, volume expiratório forçado em 1 segunda e capacidade vital (GUYATT *et al.*, 1985).

Outros estudos observaram que o *Short Physical Performance Battery* (*SPPB*), que avalia as limitações funcionais dos membros inferiores, inclui testes cronometrados de equilíbrio em pé, velocidade de caminhada e movimentos repetidos de cadeira, talvez o mais estudado e utilizado na pratica clinica com esta população, foi a medida com classificações mais positivas no que tange as propriedades psicométricas (confiabilidade teste-reteste, validade, responsividade e sensibilidade) entre outras medidas de desempenho físico (FREIBERGER *et al.*, 2012; OSTIR *et al.*, 2002).

Há evidências crescentes de que medidas objetivas de desempenho físico, como força de preensão manual, velocidade de caminhada, levantar da cadeira e equilíbrio em pé, não apenas caracterizam a capacidade física, mas também atuam como marcadores de risco a fraturas e saúde atual, permitindo a detecção precoce de distúrbios antes que surja uma perda funcional mais grave levando a incapacidades (COOPER *et al.*, 2010; ROZZINI *et al.*, 1993).

Um dos primeiros relatos de teste de desempenho para tornar mais objetiva a forma de avaliar a função de pessoas idosas, foi desenvolvido por Kruiansky e Gurland (1976), chamado de *Performance Test of Activities of Daily Living* (*PADL*) foi projetado para medir objetivamente a capacidade de autocuidado de pacientes psiquiátricos geriátricos. Após alguns anos estes testes começaram a ser validados para a população que vivia na comunidade ou residia em instituições de longa permanência, tornando mais fácil avaliar a função física clinicamente.

Dentre os testes, destaca-se a inclusão de testes que avaliam as limitações funcionais dos membros inferiores, como o teste de subir escadas (9 a 12 degraus), de caminhada (15,24m), de andar e girar 360º e o próprio *SPPB* que foram capazes de entregar informações mais confiáveis da função física do que os questionários de avaliação autorrelatados usados rotineiramente para diagnosticar essa população, pois de uma forma mais objetiva avaliava a mobilidade (GURALNIK *et al.*, 1994; REUBEN; SIU, 1990).

Destacam-se também testes que avaliam as limitações funcionais dos membros superiores, como o *Upper Extremity Physical Performance Battery* (*UEPPB*), que foi construído a partir da velocidade de caminhada de 10 pés, repetidos movimentos de sentar e levantar da cadeira e escores de equilíbrio de maneira idêntica aos estudos do *Lower Extremity Physical Performance Battery* (*LEPPB*), no entanto, inclui alguns índices com tarefas especificas de como escrever uma frase, pegar pequenos objetos, abotoar uma camisa, e alcance funcional para avaliar por exemplo: as limitações físicas autorreferidas, a incapacidade autorreferida, e dependência em *ABDV* e *AIVD* (HAZUDA *et al.*, 2005).

Embora a força de preensão, avaliada por um dinamômetro portátil, não seja uma medida pura de limitações funcionais, é um preditor robusto de incapacidade e outros desfechos clinicamente relevantes (BOHANNON, 2008).

Alguns testes foram criados e modificados para reduzir os potenciais efeitos teto e piso, que indicam os limites de variação da mudança detectável para além do qual nenhuma melhoria adicional ou deterioração pode ser percebida. Especialmente entre pessoas idosas de alto funcionamento, testes mais desafiadores de desempenho físico foram desenvolvidos para reduzir os potenciais efeito teto. Por exemplo, no *Health ABC Study*, incluiu pessoas sem deficiência de 70 a 79 anos, que avalia a velocidade de caminhada acima de 20 m, a distância percorrida em 2 minutos e o tempo para caminhar 400 m; e o *SPPB* foi modificado estendendo os tempos para os três testes padrão de equilíbrio de 10 para 30 segundos e adicionando um teste de pé unipodal (SIMONSICK *et al.*, 2001).

Outros testes foram então sendo incorporados na busca de melhor avaliar o idoso frágil e hospitalizado, como o teste de sentar e levantar e testes de equilíbrio. Estes testes agora davam uma classificação usando pontuações de apto e não apto, e em níveis que permitiam sua aplicação tanto na prática clínica quanto na pesquisa (WINOGRAD *et al.*, 1994). A seguir será apresentada um quadro dos testes com maior número de aplicações para detectar mudanças clinicamente relevantes na saúde do idoso com informações sobre a origem da validação de cada um para avaliar os domínios do desempenho físico funcional: *TUG*, TSLC, TVC e TPMI (quadro 1).

Quadro 1- Origem dos testes de desempenho físico funcional (*TUG*, TSLC, TVC e TPMI)

TESTE FUNCIONAL	ESPECIFICAÇÕES
Timed Up and Go Test (TUG) (PODSIADLO; RICHARDSON, 1991; MATHIAS; NAYAK; ISAACS, 1986).	**Objetivo:** Avaliar habilidades básicas de mobilidade em uma população de idosos frágeis da comunidade. **Procedimentos:** O indivíduo sentado com as costas contra a cadeira (altura aproximada de 46 cm) com seus braços apoiados nos braços da cadeira (altura do braço 65 cm) ele deve se levantar e caminhar em um ritmo confortável e seguro até uma linha no chão a 3 m de distância, virar, retornar à cadeira. Pode haver uma familiarização uma vez antes de ser cronometrado. Seja um relógio de pulso com ponteiro de segundos ou um cronômetro pode ser usado para cronometrar o desempenho. **Público-alvo:** A população do estudo foi composta por 60 idosos da comunidade (23 homens e 37 mulheres com idade média de 79,5 anos, variação de 60-90 anos). O teste de confiabilidade foi realizado durante um período de 2 meses. Aqueles com doença de Parkinson em estágio IV e aqueles que eram clinicamente instáveis foram excluídos do estudo de confiabilidade. Nenhum era mais do que levemente demente conforme medido pelo estado mini mental de Folstein (pontuação média 28,0). Os principais diagnósticos médicos foram acidente vascular cerebral (AVC) (n = 23), doença ou síndrome de Parkinson (n = 10), reumatoide ou osteoartrite (n = 9) e condições diversas, como fraturas de quadril pós-cirúrgicas, degeneração cerebelar e descondicionamento geral (n = 18). A validade do "Up & Go" por não ter um teste "padrão ouro" com o qual comparar essa nova medida e por isso, hipotetizaram que o escore "Up & Go" cronometrado se correlacionaria com o equilíbrio, a marcha do paciente velocidade e capacidade funcional, e, portanto, utilizou-se a Escala de Equilíbrio de Berg além do Índice de Barthel de AVD para uma estimativa clínica da capacidade do paciente sair sozinho. **Resultados:** A hipótese inicial foi suportada em que o escore do "Up & Go" cronometrado se correlacionaria com o equilíbrio, a velocidade da marcha e a capacidade funcional do paciente. O escore de tempo dos pacientes no "Up & Go" se relacionou bem aos seus escores na Escala de Equilíbrio de Berg (r = -0,72), sua velocidade de marcha (r = -0,55) e seus escores no Índice de Barthel de AVD (r = -0,51). A capacidade do escore de tempo "Up & Go" de refletir habilidades de mobilidade funcional também pode ser vista com alta relação na classificação do Índice de Barthel de AVD. A realização do teste em menos de 10 segundos indica um indivíduo totalmente independente. Todos os que completaram o "Up & Go" em menos de 20 segundos foram independentes para transferências básicas. Aqueles que demoraram mais de 30 segundos para completar o teste, por outro lado, tenderam a ser muito mais dependentes.

Quadro 1- Cont.

TESTE FUNCIONAL	ESPECIFICAÇÕES
Chair Stands Test (CSUKA; MCCARTY, 1985)	**Objetivo:** Avaliar através de um método simples, rápido e reprodutível a quantificação da força muscular de membros inferiores de pacientes com miopatia. **Procedimentos:** Foi utilizada uma cadeira de encosto reto moldado em plástico com 44,5 cm de altura e 38 cm de profundidade. O tempo necessário para completar 10 posições completas a partir da posição sentada foi registrado com um cronômetro até o décimo de segundo mais próximo. Um stand de prática foi realizado para posicionamento e aprendizado da tarefa. Sujeitos e pacientes foram então encorajados a realizar a tarefa o mais rápido possível. Todas as posições foram realizadas com os pés descalços ou com um sapato de salto baixo. O uso simultâneo das extremidades superiores não foi permitido. **Público-alvo:** Para fins de padronização, 139 indivíduos saudáveis (77 homens, 62 mulheres) realizaram o teste. Suas idades variaram de 20 a 65 anos. As pessoas foram excluídas da análise se tivessem artrite sintomática, doença pulmonar obstrutiva grave, angina sintomática, insuficiência cardíaca congestiva descompensada, obesidade mórbida ou outra doença sistêmica que impedisse sua capacidade de desempenho. A altura e o peso de cada sujeito foram registrados. Pacientes. O teste padronizado foi usado para avaliar e acompanhar seis pacientes consecutivos que se apresentaram na clínica com polimiosite tipo I (três) e tipo II (três). O diagnóstico foi confirmado em cada caso de polimiosite por níveis elevados de creatina quinase sérica, achados eletromiográficos tipicamente anormais e um resultado positivo de biópsia muscular. A função tireoidiana foi normal em cada caso. Os valores normais de creatina quinase em nosso laboratório são de 30 a 175 Ill/litro para mulheres e 30 a 200 Ill/litro para homens. **Resultados:** O peso foi relacionado ao tempo ($p<0,05$) após o ajuste para idade, mas o aumento de r^2 além daquele para idade foi de apenas 0,03, um aumento de 0,50 para 0,53. As equações de predição foram: Mulheres: Tempo (segundos) = 7,6 i- 0,17 X idade Homens: Tempo (segundos) = 4,9 + 0,19 X idade. As correlações entre idade e tempo foram de 0,7 1 para mulheres e 0,88 para homens; ambos foram estatisticamente significativos ($p <0,001$). A inclinação de 0,167 para as mulheres não foi significativamente diferente da inclinação de 0,195 para os homens. No entanto, os interceptos de 7,58 (mulheres) e 4,92 (homens) foram significativamente diferentes ($p<0,05$).

Quadro 1- Cont.

TESTE FUNCIONAL	ESPECIFICAÇÕES
Teste de Velocidade de Caminhada (BUCHNER *et al.*, 1996)	**Objetivo:** Testar uma relação não linear entre força e velocidade da marcha. O estudo também avaliou se um limiar pode ser identificado no qual a perda de força relacionada à idade começa a afetar a velocidade da marcha e a utilidade das medidas de força relativa. **Procedimentos:** A força das pernas foi medida em quatro grupos musculares (extensor do joelho, flexor do joelho, flexor plantar do tornozelo, dorsiflexor do tornozelo) de ambas as pernas usando um isocinético dinamômetro. Um escore de força da perna foi calculado como a soma das quatro medidas de força na perna direita. A velocidade usual da marcha foi medida em um percurso de 15,2 m. Os sujeitos iniciaram o teste em pé e foram instruídos a caminhar em seu ritmo habitual. A força foi medida em ambas as pernas, com velocidade de rotação da articulação do joelho de 60°/s e velocidade de rotação da articulação do tornozelo de 30°/s. Os sujeitos foram familiarizados com o protocolo antes do teste. **Público-alvo:** Uma amostra populacional de adultos com idade entre 60-96 anos (n = 409), Os adultos foram excluídos se tivessem: (1) condições neurológicas afetando o músculo esquelético (por exemplo, acidente vascular cerebral, poliomielite, demência); (2) doenças musculoesqueléticas que afetam os músculos (por exemplo, polimialgia reumática, artrite reumatoide); (3) doença sistêmica com efeitos no músculo (por exemplo, hipertireoidismo, uso crônico de corticosteroides); (4) incapacidade de andar ou doença terminal. Adultos com patologias não musculares influenciando a caminhada (por exemplo, artrite do joelho) não foram excluídos. **Resultados:** A confiabilidade das medidas de força da perna foi excelente (por exemplo, a correlação teste-reteste no mesmo dia foi Pearson r = 0,95 para a força do extensor do joelho). A confiabilidade das medidas de marcha, isto é, teste-reteste no mesmo dia foi Pearson r = 0,94. As oito medidas de força das pernas foram altamente correlacionadas. As correlações entre as pernas esquerda e direita no mesmo grupo muscular foram altas (R = 0,80-0,89), e ligeiramente superiores às correlações entre os diferentes grupos musculares da mesma perna (R = 0,67-0,87). As cinco medidas resumidas da força absoluta das pernas provaram ser quase idênticas. A correlação entre os escores sumários foi R = 0,97-0,99. A correlação entre força/peso e força/peso/altura foi extremamente alto (R = 0,98). Uma relação não linear entre a força das pernas e a velocidade da marcha que é semelhante para homens e mulheres mais velhos. Essa descoberta representa um mecanismo de como pequenas mudanças na capacidade fisiológica podem ter efeitos substanciais no desempenho em adultos frágeis, enquanto grandes mudanças na capacidade têm pouco ou nenhum efeito em adultos saudáveis.

Quadro 1- Cont.

TESTE FUNCIONAL	ESPECIFICAÇÕES
Teste de Potência para Membros Inferiores (LINDEMANN *et al.*, 2003)	**Objetivo:** Validar um método seguro e portátil para medir a potência de saída na transferência do teste de sentar e levantar, representando uma tarefa funcional. Além disso, tenta comparar os resultados com outros métodos comumente usados. **Procedimentos:** Os sujeitos sentaram-se em uma cadeira de altura padrão (46 cm) com os braços cruzados sobre o peito, cada pé apoiado sem restrições em uma plataforma de força e seu tronco tocando o encosto da cadeira. Eles foram instruídos a se levantarem o mais rápido possível. Assim, a potência média foi calculada a partir da força vertical do peso corporal, a diferença entre a altura na posição sentada e na ereta e o tempo necessário para levantar-se (T2-T3). Foi calculado a partir da seguinte equação: p= f x s x t -¹ ð1Þ onde os fatores constantes peso corporal (f) e distância (s) representam a força de reação do solo do peso corporal e a diferença entre a altura do corpo sentado e em pé posição, respectivamente. O intervalo de tempo do PR (T2–T3) foi expresso como tempo (t). **Público-alvo:** Para uma avaliação transversal uma amostra de conveniência de 33 idosos saudáveis [média (DP): 67,8 (6,7) anos; 17 homens, 16 mulheres] foram recrutados. Os indivíduos não tinham problemas neurológicos ou ortopédicos conhecidos e foram sucessivamente inscritos. Todos os indivíduos deram consentimento informado por escrito. **Resultados:** A confiabilidade da medição repetida foi avaliada em 31 jovens funcionários saudáveis do hospital [37,9 (7,6) anos]. Duas séries no mesmo dia, tomando a melhor de duas tentativas, de acordo com os resultados de potência, foram comparadas para descrever a confiabilidade teste-reteste (r icc=0,95). Foi dado um intervalo de 4 minutos entre as séries. A confiabilidade entre avaliadores foi avaliada comparando-se a primeira série com o mesmo protocolo no dia seguinte, avaliada por um investigador diferente (r icc=0,96). O tempo médio da transferência total de STS (T1–T4) foi de 1,2 (0,07) s para os idosos saudáveis. De acordo com as diferentes fases durante a transferência descritas acima, o tempo de sentar-se para as pernas estendidas (RP, T2-T3) levou 37,5% [0,45 (0,07) s] do movimento total. PP (T1-T2) e SP (T3-T4) representaram 52,5% [0,63 (0,11) s] e 10% [0,12 (0,08) s], respectivamente, do movimento total. A medição da potência durante a transferência de sentar para levantar mostrou boa correlação com a medição da força isocinética (r=0,68) e com o ‘‘‘Nottingham power rig’’ (r=0,6). A correlação com a elevação de cinco cadeiras foi ruim (r=)0,08). Em conclusão, o estudo mostra que o método apresentado é capaz de medir a potência durante a execução de uma tarefa diária. A fraca correlação entre a medida introduzida e a elevação de cinco cadeiras sugere que pode ser capaz de detectar declínio na função muscular mais cedo pela medida introduzida do que pela medida do estado funcional.

Fonte: Elaborado pelo autor.

3 MATERIAIS E MÉTODOS

3.1 Protocolo e registro

Esta revisão sistemática foi escrita de acordo com as diretrizes *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (*PRISMA*) (ARDERN *et al.*, 2022). O protocolo foi registrado no *International Prospective Register of Systematic Reviews* (*PROSPERO*) (CRD42021239084) (BOOTH *et al.*, 2012).

O Grupo *COSMIN* (*COnsensus-based Standards for the selection of health Measurement INstruments*) desenvolveu uma diretriz para a realização de revisões sistemáticas sobre *PROMs* (*Patient-Reported Outcome Measures*) e forneceu ferramentas para avaliar a qualidade de suas propriedades de medição. Essas diretrizes foram seguidas ao longo desta revisão sistemática bem como foi o protocolo (MOKKINK, *et al.*, 2018; PRINSEN, *et al.*, 2018; TERWEE, *et al.*, 2018).

3.2 Taxonomia *COSMIN*

A iniciativa *COSMIN* desenvolveu uma taxonomia de propriedades de medição relevantes para avaliar um *PROM* (MOKKINK *et al.*, 2010b). Ao avaliar a qualidade de um *PROM*, distingue-se três domínios: confiabilidade, validade e capacidade de resposta. O domínio confiabilidade contém três propriedades de medição: consistência interna, confiabilidade e erro de medição. O domínio validade também contém três propriedades de medida: validade de conteúdo, teste de hipóteses para validade de construto e validade de critério. Para a presente revisão será adotado apenas a avaliação das seguintes propriedades: confiabilidade, erro de medição, validade de critério e validade de construto.

3.2.1 Confiabilidade

A confiabilidade que enquanto domínio, é a capacidade de um instrumento de avaliação obter repetidamente a mesma medida na ausência de mudanças reais. Por exemplo, a utilização de diferentes conjuntos de itens em um mesmo instrumento (propriedade de medição consistência interna); ao longo do tempo (propriedade de medição teste-reteste); por pessoas diferentes na mesma ocasião (propriedade de medição entre avaliadores); ou pelas mesmas pessoas, ou seja, avaliadores ou avaliados, em ocasiões diferentes (propriedade de medição intra avaliador); a proporção da variância total nas medidas que se deve as diferenças “verdadeiras” entre avaliados (propriedade de medição confiabilidade); erro sistemático e aleatório da pontuação de um avaliado que não é atribuído a mudanças verdadeiras no construto a ser medido (propriedade de medição erro de medição do domínio confiabilidade).

3.2.2 Validade

O domínio validade, que é o grau em que um instrumento de medida mede o(s) construto(s) que pretende medir. Por exemplo, o grau em que as pontuações de um instrumento são consistentes com as hipóteses (propriedade de medição validade de construto e validade estrutural, transcultural e teste de hipóteses como aspectos desta propriedade); O grau em que as pontuações de um instrumento são um reflexo adequado de um "padrão ouro" (propriedade de medição validade de critério) (MOKKINK *et al.*, 2010b).

3.3 Estratégia de pesquisa

A procura literária foi realizada usando as bases de dados eletrônicas Biblioteca Virtual em Saúde (BVS), *PubMed e Embase* com o intuito de buscar publicações que englobam desde a américa latina e caribe até os estados unidos da américa e a europa durante o mês de novembro de 2021 resultando em 693 estudos totais, e uma nova busca no mês de janeiro de 2023 resultando em 88 estudos totais. A busca na literatura foi realizada por um pesquisador e não houve restrição quanto à idioma e data de publicação. Para a estratégia de busca, empregou-se a combinação dos termos *MeSH* junto aos termos que mais são utilizados na literatura para busca metodológica altamente sensível (97,4%) para encontrar estudos sobre propriedades de medição no *PubMed* (apêndice A), posteriormente adaptado para as outras bases de dados [1. Construto ou fenômeno de interesse; 2. População de interesse; 3. Instrumento de interesse; 4. 1 *AND* 2 *AND* 3 *AND* incluir as propriedades de mensuração de interesse; 5. 4 *NOT* (filtro de exclusão)] (TERWEE *et al.*, 2009).

As seguintes palavras-chave foram usadas para capturar o termo desempenho físico funcional: *functional performance physical, functional performances physical, functional performances physical, performance physical functional, performances physical functional, physical functional performances, functional performance, functional performances, performance functional, performances functional, physical performance, performance physical, performances physical, physical performances*. *As seguintes palavras-chave foram usadas para capturar o termo idoso: elderly, aging, age, aged, elder, elders, older, old, sênior, geriatric, oldest old, older people, older adults, frail elderly, frail elder, frail older, functionally impaired elderly, elderly functionally impaired, very elderly*.

As seguintes palavras-chave foram usadas para capturar os testes de desempenho funcional físico (*TUG*): *timed up and go test, timed up go test, timed, up & go, timed up & go, get up and go test, TUG, GUG, TGUG, go, TGUGT, TUGT, modified TUG, ITUG*.

As seguintes palavras-chave foram usadas para capturar os testes de desempenho funcional físico (TSLC): *sit-to-stand test, sit-to-stand, stand-to-sit, chair stand, STS, 5STS, 30STS, five-repetition sit-to-stand test, 5 times sit-to-stand test.*

As seguintes palavras-chave foram usadas para capturar os testes de desempenho funcional físico (TVC)*: gait speed, walking speed, walking speed, gait velocity test, 10 meter walking test, 6MWT, 10MWT, 6 meter walking test, 5 meter walking test, 4 meter walking test, gait velocity*.

As seguintes palavras-chave foram usadas para capturar os testes de desempenho funcional físico (TPMI): *vertical jump test, power test, countermovement, countermovements, countermovement jump test, squat jump, squat jump test, agility test, change of Direction, change of direction test.*

As seguintes palavras-chave foram usadas para capturar as propriedades psicométricas e propriedades de medição: *psychometrical, psychometrically psychometrics, psychometrics, psychometric, measurability, measurable, measurably, measures, measureable, measured, measurement, measurements, measurer, measurers, measuring, measurings, measurment, measurments, weights and measures, weights, measures, weights and measures, measure, measures, reliabilities, reliability, reliable, reliablity, reliably, repeatabilities, repeatability, repeatable, repeated, repeatibility, reproducability, reproducibilities, reproducibility, producible, measurement error, consistence, consistences, consistencies, consistency, consistente, consistently, smallest worthwhile change, minimal detectable change, typical error, useful, usefulness, valid, validate, validated, validates, validating, validation, validational, validations, validator, validators, validities, validity, logic, logic, logics, logical, logically, constructs, constructed, constructing, construction, constructions, constructive, constructively, constructs, constructo, converge, converged, convergence, convergences, convergencies, convergency, convergente, convergently, convergentes, converges, converging, discriminabilities, discriminability, discriminable, discriminably, discriminance, discriminant, discriminants, discriminate, discriminated, discriminates, discriminating, discrimination, psychological, discrimination, psychological, psychological discrimination, discrimination, discriminations, discriminative, discriminatively, discriminator, discriminators, gold standard, level, levels, reference standards, reference, standards, reference standards, standardization, standard, standards, standardisation, standardisations, standardise, standardised, standardises, standardising, standardizations, standardizations, standardize, standardized, standardizes, standardizing, standards*.

3.4 Critérios de elegibilidade

3.4.1 Critérios de inclusão

Os critérios de elegibilidade para inclusão no estudo consistiram em um dos seguintes: (i) estudos que utilizaram qualquer um dos seguintes testes e suas variações, variações estas que não comprometessem a praticidade e acessibilidade dos mesmos para avaliar o desempenho físico funcional: *TUG*, TSLC, TVC e TPMI; (ii) investigar e descrever de forma clara pelo menos uma das propriedades de medição e/ou domínios mencionadas pela iniciativa *COSMIN* para boas propriedades de medição dentro dos domínios que a revisão se propõe a investigar; (iii) incluir populações com 60 anos ou mais considerados institucionalizados ou da comunidade capazes deambular independentemente com ou sem um dispositivo auxiliar; (iv) ter um tamanho de amostra de ≥ 30 participantes.

3.4.2 Critérios de exclusão

Os estudos foram excluídos quando o desempenho físico funcional foi avaliado por meio de instrumentos que avaliam vários domínios físicos funcionais como as baterias de testes ou se o instrumento utilizado foi desenvolvido para populações com doenças específicas (por exemplo, pacientes com doença de Parkinson, Alzheimer ou qualquer doença neurológica com déficits motores, condição terminal, doença ou lesão aguda, como infecção ou inflamação e dor limitante da atividade, ou seja, testes para grupos de pacientes). Com exceção das revisões, todos os tipos de estudos de proposição e/ou validação foram relacionados, e todos foram publicados em um periódico revisado por pares sem restrição de gênero (feminino e masculino) e de idioma.

3.5 Seleção da literatura

A seleção da literatura consistiu em três fases de triagem por dois revisores independentes utilizando o aplicativo web móvel para revisões sistemáticas *Ryyan* (OUZZANI *et al.*, 2016). Na primeira fase, as duplicatas foram selecionadas e removidas; na fase dois, títulos e resumos foram analisados e selecionados os artigos a partir desta análise; na fase três, os artigos selecionados na fase dois, foram lidos na íntegra para uma análise mais minuciosa usando os critérios de elegibilidade (inclusão) já mencionados. Qualquer discordância para a inclusão entre os revisores, foi decidida por um terceiro revisor, em todos os estudos selecionados com base nos critérios de inclusão para a seleção dos estudos que foram incluídos para leitura do texto completo.

3.6 Extração dos dados

Os dados extraídos de cada estudo selecionado corresponderam à estratégia abordada na sessão 3.3 que incluiu: o teste realizado, o construto a ser medido, população de interesse, além de detalhes da publicação como número de participantes; informações demográficas (incluindo sexo, idade, país e idioma); nome do teste; breve descrição do teste; medidas de resultado bem como as propriedades de medição dos testes e informações necessárias para avaliar a qualidade metodológica dos estudos e apresentação das características dos estudos incluídos (tabela 1). Os dados foram extraídos de forma independente por dois revisores e documentados em planilha *Microsoft Excel* 2019 (*Microsoft Corporation, Redmond, Washington, EUA*) seguido da posterior análise de imparcialidade dos dois primeiros pesquisadores por um terceiro revisor.

3.7 Avaliação da Qualidade Metodológica dos Estudos

3.7.1 *COSMIN Risk of Bias Checklist*

A avaliação da qualidade metodológica dos estudos incluídos foi realizada utilizando o *COSMIN Risk of Bias Checklist* que possui doze itens separados por caixas nas quais dez são usadas para avaliar se o estudo satisfaz as normas de boa qualidade (MOKKINK *et al.*, 2018).

As propriedades de medição avaliadas em um artigo determinam quais caixas precisam ser preenchidas. Por exemplo, como foi avaliado somente a confiabilidade, erro de medição, validade de critério e teste de hipóteses para validade de construto, apenas quatro caixas precisaram ser preenchidas, ou seja, caixa 6, 7, 8 e 9 (apêndice E) respectivamente. Este sistema modular foi desenvolvido porque nem todas as propriedades de medição são avaliadas em todos os artigos, como os testes a serem pesquisados nesta revisão, por se tratar de testes físicos com medidas objetivas, é muito comum utilizarem apenas as propriedades de medição supracitadas.

Para o risco de viés foi desenvolvido um sistema de classificação qualitativo de quatro pontos classificando como "*very good*/muito bom, *adequate*/adequado, *doubtful*/duvidoso ou *inadequate*/inadequado". Os critérios abrangem, por exemplo, tratamento de itens ausentes, tamanho da amostra e adequação dos métodos estatísticos. A classificação geral da qualidade de cada estudo individual em uma propriedade de medição, é obtida considerando-se a pior classificação alcançada para um dos itens específicos de cada propriedade. Por exemplo, se a classificação mais baixa dos nove itens da caixa confiabilidade for "inadequada", a qualidade metodológica geral desse estudo é classificada como "inadequada". Para avaliar a qualidade dos estudos sobre confiabilidade e erro de medição, foi utilizada a ferramenta *COSMIN Risk of*

Bias específica para este domínio e sua respectiva propriedade de medição (MOKKINK *et al.*, 2020).

3.7.2 Critérios de avaliação para boas propriedades de medição

Subsequentemente, o resultado de cada estudo foi classificado em relação aos critérios atualizados para boas propriedades de medição, onde cada resultado pôde ser classificado como suficiente (+), insuficiente (-) ou indeterminado (?). Para avaliar a qualidade metodológica dos estudos de confiabilidade e erro de medição, consideramos intervalo de correlação intraclasse (*ICC*), kappa ponderado e concordância como medidas adequadas de confiabilidade e limites de concordância (*LoA*), menor alteração detectável (*SDC*) e alteração mínima importante (*MIC*) como medidas adequadas de erro de medição. Os coeficientes de correlação de Pearson e Spearman são considerados menos adequados, pois negligenciam erros sistemáticos entre as medidas (MOKKINK, *et al.*, 2020, p 37; WEIR, 2005). Por isso, as correlações de Pearson e Spearman não foram levadas em consideração nesta revisão.

Sendo assim, um instrumento recebeu pontuação "+" quando possuía alta confiabilidade [*ICC* ou *Kappa* ponderado ≥ 0,70; para erro de medição, a pontuação "+" é empregada quando a *SDC* ou *LoA* foram menores que a *MIC*; deu-se uma classificação "+" para a validade de construto quando as hipóteses foram especificadas antecipadamente e quando pelo menos 75% dos resultados estivam em correspondência com essas hipóteses, em subgrupos de pelo menos 50 pacientes, ou seja, alta validade de construto quando correlação entre construtos for ≥ 0,50; para a validade de critério, empregou-se a pontuação "+" se quando apresentados argumentos convincentes de que o padrão usado realmente é "ouro" e quando a correlação com o padrão-ouro foi de pelo menos 0,70, ou seja, alta validade de critério quando correlação de Pearson/Spearman ou área sob a curva (*AUC*) forem ≥ 0,70] (PRINSEN *et al.*, 2016; TERWEE *et al.*, 2007).

3.7.3 Sumarização dos resultados dos estudos por teste

Posteriormente, todos os estudos relacionados ao mesmo teste foram comparados quanto à qualidade geral, consistência nos achados e agrupados quantitativamente e comparados com os critérios para boas propriedades de medição para determinar se, em geral, a propriedade de medição do teste é suficiente (+), insuficiente (-), inconsistente (±) ou indeterminado (?). Se os resultados por estudo fossem todos suficientes ou insuficientes, a classificação geral também foi suficiente ou insuficiente. Em caso de mais de um estudo avaliando a mesma propriedade, se eles apresentassem inconsistência entre si, os estudos com

maior qualidade tiveram um maior peso na sumarização, assim como os estudos mais recentes em comparação com os mais antigos (MOKKINK *et al.*, 2018).

3.7.4 Graduação da qualidade da evidência

Uma abordagem *GRADE* (*Grading of Recommendations Assessment, Development and Evaluation*) foi empregada para classificar a qualidade geral graduando o nível de certeza das evidências para propriedades de medição de cada estudo (GUYATT *et al.*, 2011). Especificamente, foi aplicado uma abordagem modificada, conforme recomendado pela iniciativa *COSMIN*, para a classificação da evidência *very low quality, low quality, moderate or high*. (PRINSEN *et al.*, 2018).

Para avaliar as propriedades de medição em revisões sistemáticas de *PROMs,* os quatro fatores a seguir são levados em consideração: (1) risco de viés (ou seja, a qualidade metodológica dos estudos), (2) inconsistência (ou seja, inconsistência inexplicada dos resultados entre os estudos), (3) imprecisão (ou seja, tamanho total da amostra dos estudos disponíveis) e (4) indiretividade (ou seja, evidências de populações diferentes da população de interesse na revisão). O quinto fator, ou seja, viés de publicação, é difícil de avaliar em estudos sobre propriedades de medida, devido à falta de registros para esse tipo de estudo. Portanto, esse fator não é levado em consideração nesta metodologia.

A abordagem *GRADE* é usada para rebaixar a evidência quando há incertezas sobre a qualidade da evidência. O ponto de partida é sempre a suposição de que o resultado combinado ou geral é de alta qualidade. A qualidade da evidência é posteriormente rebaixada em um ou dois níveis por fator para evidência moderada, baixa ou muito baixa quando há risco de viés, inconsistência (inexplicável), imprecisão (baixo tamanho da amostra) ou resultados indiretos. A qualidade da evidência pode até ser rebaixada em três níveis quando a evidência é baseada em apenas um estudo inadequado (ou seja, risco extremamente sério de viés).

Todas as etapas citadas acima (avaliação da qualidade metodológica dos estudos individuais, avaliação da qualidade das propriedades de medição individual e geral e avaliação da qualidade da evidência) foram realizadas por dois revisores de forma independente, nas divergências quando ambos mantiveram suas decisões, um terceiro revisor foi consultado para que então o consenso fosse alcançado.

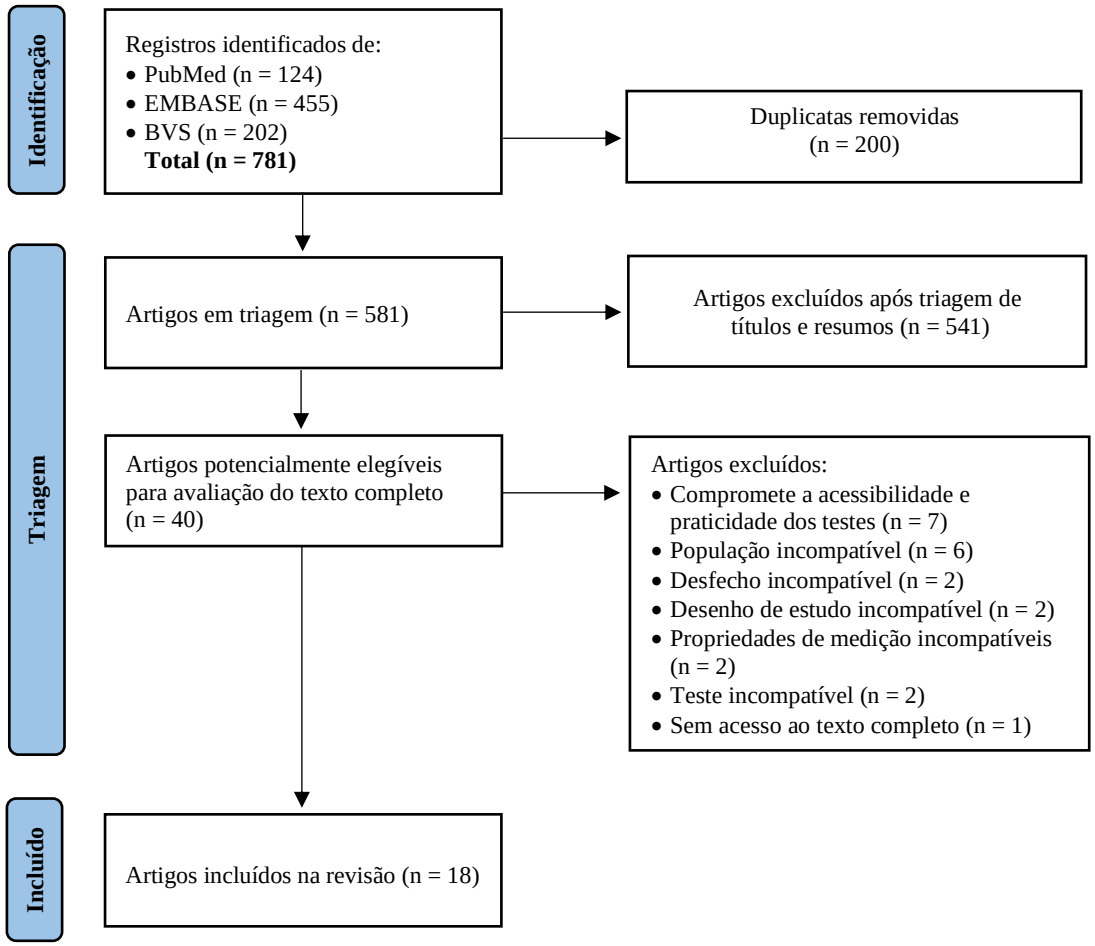
4 RESULTADOS

4.1 Resultados da estratégia de pesquisa

A partir da estratégia de busca, foram encontrados 781 estudos potencialmente relevantes. Destes, apenas 18 estudos envolvendo um total de 1.204 pessoas idosas (mulheres= 58% e homens= 42%) com 1.077 (89%) sendo pessoas idosas da comunidade e 127 (11%) pessoas idosas institucionalizadas, foram considerados elegíveis para análise de dados. Dos 18 estudos elegíveis, 8 estudos corresponderam ao teste *TUG*, 3 estudos o TSLC, 9 estudos o TVC e quatro estudos o TPMI. Alguns estudos constaram para mais de um teste.

Houve grandes diferenças nos tamanhos das amostras de cada estudo, variando de 32 a 136 participantes. Em 15 estudos foi avaliado o domínio confiabilidade, 13 estudos avaliaram a propriedade de medição erro de medição, 2 estudos avaliaram o teste de hipóteses aspecto da propriedade de medição validade de construto e 8 estudos avaliaram a propriedade de medição validade de critério. Alguns estudos avaliaram um, dois ou três domínios, mas nenhum avaliou todos os quatro domínios analisados na revisão. Uma visão geral completa do processo de triagem pode ser encontrada na figura 1.

Figura 1 - Diagrama de fluxo *PRISMA* para triagem de artigos



4.2 Características dos estudos elegíveis

Dos oito estudos correspondentes ao *TUG*, cinco são com pessoas idosas da comunidade (CHAN *et al.*, 2016; COLLADO-MATEO *et al.*, 2019; DEWHURST; BAMPOURAS, 2014; GINÉ-GARRIGA *et al.*, 2010; LEE *et al.*, 2016) e três são com pessoas idosas institucionalizadas (GALHARDAS; RAIMUNDO; MARMELEIRA, 2020; LE BERRE *et al.*, 2016; NEPAL; BASAULA; SHARMA, 2020) totalizando 529 pessoas idosas (m= 314 e h= 215) com média de idade que varia de 63.3 a 91.3 anos. Destes oitos estudos, houve a utilização do *TUG* cronometrado por cronômetro digital portátil, aplicativo de smartphone, e por sensor de força laboratorial.

Dos três estudos que correspondem ao TSLC, dois estudos são com pessoas idosas da comunidade (CHAN *et al.*, 2016; COLLADO-MATEO *et al.*, 2019) e um com participantes pessoas idosas institucionalizadas (LE BERRE *et al.*, 2016) totalizando 185 pessoas idosas (m= 54 e h= 131) com idade média que varia de 70.7 a 91.3 anos. Estes três estudos, envolveram a utilização do TSLC de 30 segundos, de 5 repetições e um modificado de 30 segundos com o uso auxiliar dos membros superiores cronometrado por cronômetro digital portátil, aplicativo de smartphone e por sensor de força laboratorial.

Todos os nove estudos correspondentes ao TVC foram com pessoas idosas da comunidade (ADELL; WEHMHÖRNER; RYDWIK, 2013; CRISS *et al.*, 2023; DEWHURST; BAMPOURAS, 2014; FERNÁNDEZ-HUERTA; CÓRDOVA-LEÓN, 2019; FORTE; DE VITO; BOREHAM, 2021; GINÉ-GARRIGA *et al.*, 2010; ÖZDEN *et al.*, 2022; SAITO *et al.*, 2022; SAYERS *et al.*, 2006) totalizando 645 pessoas idosas (m= 401 e h= 244) com média de idade que varia de 68.9 a 89 anos. Estes nove estudos, envolveram a utilização do TVC com distâncias de 3, 4, 6, 7, 8, 10, 12.24 e 400 m com velocidade de execução auto selecionada ou habitual, lenta e máxima, cainhando para trás e para frente, com tarefa única e dupla tarefa (cognitiva e motora), não cronometrado e cronometrado por cronômetro digital portátil e aplicativo de smartphone.

Todos os quatro estudos do TPMI (BALACHANDRAN *et al.*, 2021; GINÉ-GARRIGA *et al.*, 2010; SHERWOOD *et al.*, 2020) são com pessoas idosas da comunidade, com exceção do estudo de (CRUVINEL-CABRAL *et al.*, 2018), que sua amostra é dividida por pessoas idosas da comunidade institucionalizadas totalizando 177 (m= 134 e h= 43) com média de idade variando entre 69.4 e 77.6 anos. Estes quatro estudos, envolveram o TSLC de 5 e 1 repetição máxima, salto contramovimento, força de levantar da cadeira e chutar uma bola medidos por *LPT*, aplicativo de smartphone e cronômetro digital portátil. Uma visão geral completa dos artigos incluídos na revisão pode ser encontrada na tabela 1.

Tabela 1 - Características dos estudos elegíveis

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
TUG						
Collado-Mateo *et al.*, (2019)	TUG com um cronômetro automático foi colocado na cadeira para avaliar o tempo necessário para completar a tarefa. Especificamente, o cronômetro foi o Chronopic (Chronojump, BoscoSystem®, Barcelona, Espanha). O principal objetivo deste estudo é fornecer parâmetros de confiabilidade para o TUG e comparar os resultados registrados com cronômetro manual stopwatch e cronômetro automático no TUG, bem como relatar a confiabilidade, erro padrão de medição e menor diferença real.	99 (M= 33 e H= 66)	71.10 (6.02)	Da comunidade	Função física e parâmetros cinemáticos	[a]TUG cronômetro automático Confiabilidade (ensaio 2) de 5 ensaios de teste-reteste: ICC= 0.892 (95% CI= 0.843–0.926) SEM=0.55/5.53% SRD= 1.51/15.32% Tempo do reteste: 1 minuto [b]TUG cronômetro manual Confiabilidade (ensaio 2) de 5 ensaios de teste-reteste: ICC= 0.878 (95% CI= 0.825–0.916) SEM= 0.57/5.85% SRD= 1.59/16.21% [c]Correlação entre TUG Manual vs TUG Cronometrado (Rho de Spearman's ou r Pearson's): Ensaio 1= 0.979
Galhardas *et al.*, (2020)	Todos deveriam se levantar de uma cadeira, caminhar 3 m, contornar um cone, voltar para a cadeira e voltar para a cadeira. sente-se novamente como o teste original proposto por Podsiadlo e Richardson (1991). O objetivo principal deste estudo foi examinar a confiabilidade teste-reteste de quatro testes motores de campo, dentre eles o TUG para avaliação do equilíbrio.	53 (M= 41 e H= 12)	85.9 (3.9)	Da instituição	Mobilidade física (velocidade, agilidade e equilíbrio dinâmico)	Confiabilidade teste-reteste relativa: ICC= 0.99 (95% CI= 0.99-1.00) Confiabilidade teste-reteste absoluta: SEM= 0.5 MDC 95%= 1.5 Tempo do reteste: Duas sessões (com intervalo de 10 e 14 dias)

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
Nepal *et al*., (2020)	Uma cadeira padrão, passarela com comprimento superior a 3 m, marcador de ponto de viragem, cronômetro para registro do tempo em segundos. A altura do assento da cadeira foi mantida próxima de 46 cm. O objetivo principal deste estudo foi avaliar a confiabilidade entre avaliadores de cuidadores que avaliam o TUG em idosos em comparação com a avaliação realizada por um estudante fisioterapeuta em uma comunidade rural no Nepal.	100 (M= 54 e H= 46)	69.10 (7.95)	Da instituição	Equilíbrio e mobilidade funcional	Confiabilidade entre avaliadores (estudante vs cuidadores) Pontuação média do TUG (s): Estudante= 14.8 (6.0) Cuidador= 14.8 (5.8) ICC= 0.87 (95% CI= 0.82-0.91) Tempo do reteste: 5 minutos
Chan *et al.*, (2016)	Uma cadeira com os braços a 46 cm de altura, caminhar 3 m até uma marca no chão, virar-se, retornar à cadeira e sentar-se. A tarefa foi executada em uma velocidade de caminhada confortável e individualizada. O objetivo deste estudo foi examinar a consistência e a concordância da medição de um aplicativo de smartphone recém-criado (O algoritmo do aplicativo para smartphone foi baseado nos dados coletados da unidade de medição inercial tridimensional (IMU) construída em um smartphone baseado em Android (Galaxy Note If; Samsung Electronics Co. Ltd, Suwon, Coreia). O telefone foi fixado com segurança no peito	32 (M= 21 e H= 11)	70.7 (6.5)	Da comunidade	Mobilidade funcional	Validade concorrente (Consistência da medição entre o aplicativo do smartphone e o sensor de força): ICC= 0.946 (95% CI= 0.889-0.973) Bland e Altman: Viés positivo= 0.48 s LoA 95%= -1.66 s a 2,63 s

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	do participante por tiras de velcro durante o teste.) em relação a uma condição de referência baseada em laboratório (sensor de força (YZC-516; Guangzhou Electrical Measuring Instruments Factory, Guangzhou, China) foi instalado no encosto da cadeira de teste.					
Le Berre *et al.*, (2016)	As instruções padrão foram as seguintes: "Quando eu disser 1, 2, 3 vai', por favor, levante-se da cadeira, caminhe (com seu dispositivo auxiliar) em um ritmo confortável até a fita, volte para a cadeira e sente-se." A linha ficava a 3 m da cadeira e marcada com fita adesiva no chão. Os participantes foram cronometrados com precisão de 0,1 s usando o cronômetro digital. O objetivo do estudo foi avaliar a confiabilidade teste-reteste e a validade convergente do TUG com um protocolo 30STS com componente modificado que permite o uso de membros superiores em idosos institucionalizados.	54 homens	91 (3)	Da instituição	Desempenho da função físico (velocidade e equilíbrio)	Confiabilidade teste-reteste: ICC= 0.85 (95% CI= 0.76-0.91) SEM= 3.91 (95% CI= 3.29-4.83) MDC= 9.08 Tempo do reteste: Duas sessões com intervalo previsto entre as sessões de teste de 5 dias. No entanto, isso não foi possível para alguns participantes. Validade convergente teste-reteste (TUG vs 30STS): Ensaio 1= (r=-62, (95% CI= -0.76 a -43) Ensaio 2= (r=-0.62, (95% CI= -0.76 a -0.43)
Lee *et al.*, (2016)	A altura padrão do assento de 46 cm foi usada para a primeira fase do teste. O investigador sinalizou o Início do teste dizendo "Pronto? Vá." Ao "ir", o	83 (M=57 e H= 26)	69.3 (6.9)	Da comunidade	Mobilidade funcional	Cronômetro portátil [a]3m: ICC= 0.887 (95% CI= 0.675-0.962) SEM= 0.87 s

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	participante foi instruído a levantar-se do banco, caminhar a distância predeterminada em direção a um alvo em um ritmo confortavelmente rápido, virar-se, caminhar de volta em direção ao banco, virar-se novamente e sentar-se. O objetivo do estudo foi avaliar 3 fatores procedimentais: (1) método de cronometragem (cronômetro portátil vs. cronometragem baseada em carga); (2) distância percorrida (3 m, 6 m ou 9 m): (3) e altura do assento (padronizado vs. específico individual).					MDD 95%= 2.41 s [c]6m: ICC= 0.950 (95% CI= 0.850-0.983) SEM= 0.84 s MDD 95%= 2.33 s [e]9m: ICC= 0.960 (95% CI= 0.881-0.986) SEM= 0.96 s MDD 95%= 2.66 s Tempo baseado em carga [b]3m: ICC= 0.985 (95% CI= 0.956-0.995) SEM= 0.31 s MDD 95%= 0.86 s [d]6m: ICC= 0.972 (95% CI= 0.916-0.991) SEM= 0.62 s MDD 95%= 1.72 s [f]9m: ICC= 0.968 (95% CI= 0.906-0.989) SEM= 0.84 s MDD 95%= 2.33 s
Dewhurst *et al*., (2014)	Timed 8-Foot Up-and-Go: Para avaliar a integração desses parâmetros (potência, velocidade, agilidade e equilíbrio dinâmico), tempo necessário para levantar-se da posição sentada, caminhar 2,44	71 mulheres	71.7 (7.3)	Da comunidade	Capacidade funcional para tarefas da vida cotidiana	Confiabilidade para as variáveis entre ensaios (confiabilidade teste-reteste intra sessão): 2 vs 1 ICC= 0.90 (95% CI= 0.85-0.94) 3 vs 2

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	m (8 pés), virar e retornar à posição sentada foi registrado. O objetivo do presente estudo foi avaliar a confiabilidade e sensibilidade intra sessão.					ICC= 0.93 (95% CI= 0.90-0.96) Tempo do reteste: Todos os testes foram feitos em uma única sessão com 1 minuto de descanso entre os ensaios.
Giné-Garriga *et al.*, (2010)	O teste TGUG modificado é uma ferramenta de avaliação da função física que mede o equilíbrio e a marcha com uma tarefa dupla: realizar uma tarefa cognitiva (contagem regressiva de 15 a 0) e uma tarefa física (andar em círculos) durante a caminhada: a força dos membros inferiores é medida por levantando-se de uma cadeira e chutando o mais forte possível uma bola de 19 cm, 0,2 kg, tempo total necessário para realizar o teste (TT). Os testes de comparação incluídos foram: velocidade de marcha rápida e normal (FGS e NGS), registrando o tempo que cada participante levou para caminhar os 8 m centrais de um percurso de 12 m e dividindo a distância (8 m) pelo tempo da velocidade da marcha. O objetivo foi determinar até que ponto os escores do teste TGUG modificado se correlacionaram com outras medidas comumente utilizadas na literatura para documentar o declínio da função	37 mulheres	72.3 (5.5)	Da comunidade	Desempenho físico correlacionado com declínio cognitivo (marcha com dupla tarefa)	Componentes do TGUG modificado e medidas de desempenho físico: [a]TT vs. NGS: r= 0.841 [b]TT vs. FGS: r= 0.748

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	relacionado à idade (validade de critério concorrente).					
TLSC						
Collado-Mateo *et al.*, (2019)	Chair Stand 30 s (CST30s), o cronômetro automático (Chronopic Chronojump, BoscoSystem®) foi utilizado para avaliar o tempo gasto em cada repetição. No entanto, duas fases foram identificadas no ciclo sentar-levantar-sentar-se: fase de impulso, que é definida como o tempo decorrido desde o momento em que as nádegas entram em contato com o assento até que as nádegas percam o contato com o assento (ou seja, todo o tempo em que o participante está sentado) e a fase sem contato, que é definida como o tempo decorrido desde que as nádegas perdem o contato com o assento até que o contato seja feito novamente. O principal objetivo deste estudo é fornecer parâmetros de confiabilidade para o CST30s, bem como relatar o erro padrão de medição e menor diferença real.	99 (M=33 e H=66)	71.10 (6.02)	Da comunidade	Função física e parâmetros cinemáticos	Confiabilidade teste-reteste (número de repetições): ICC= 0.874 (95% CI= 0.817-0.913) SEM= 0.66/6.82% SRD= 1.87/18.91% Tempo do reteste: 3 minutos entre os 2 ensaios
Chan *et al.*, (2016)	O teste Five-Time Sit-To-Stand (FTSTS) mediu o tempo necessário para completar cinco repetições da manobra sentar-se levantar-se o mais rápido	32 (M= 21 e H= 11)	70.7 (6.5)	Da comunidade	Mobilidade funcional	Validade concorrente (Consistência da medição entre o aplicativo do smartphone e o sensor de força YZC-516; Guangzhou Electrical

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	possível. Todos os participantes foram convidados a sentar-se em uma cadeira sem braços com 43 cm de altura. Antes do teste, os participantes cruzaram os braços sobre o peito, sentaram-se eretos e com as costas em contato com o encosto da cadeira. A manobra correta foi demonstrada e incluiu ficar em pé completo (definido como tronco ereto com quadris e joelhos estendidos). Os participantes tiveram que encostar as costas no encosto ao final de cada repetição.					Measuring Instruments Factory): ICC= 0.988 (95% CI= 0.976-0.994) Bland e Altman: Viés positivo= 0.27 s LoA 95%= -1.22 s a 1.76 s
Le Berre *et al.*, (2016)	Sit-To-Stants modificado (uso dos membros superiores) 30STS). Os participantes começaram sentados, em uma cadeira padrão (altura do assento 17 polegadas, largura do assento 18 polegadas) com apoios de braços. As instruções padrão foram as seguintes: Quando eu disser '1, 2, 3, vá', quero que você se levante e sente-se novamente. Você pode usar as mãos para ajudá-lo a ficar de pé, se necessário. Tente ficar de pé e sentar-se tantas vezes quanto possível enquanto eu cronometro você por 30 segundos. O objetivo do foi avaliar a confiabilidade teste-reteste e a validade convergente de um protocolo 30STS com	54 homens	91 (3)	Da instituição	Desempenho da função físico (velocidade e equilíbrio)	Confiabilidade teste-reteste: ICC= 0.84 (95% CI= 0.74-0.91) SEM= 1 (95% CI= 0.87-1.28) MDC= 2 Tempo do reteste: 2 ensaios com intervalo previsto entre os ensaios de teste de 5 dias. No entanto, isso não foi possível para alguns participantes. Validade convergente teste-reteste (30STS vs TUG): Ensaio 1 (r=-62, (95% CI= -0.76 a -43) Ensaio 2 (r=-0.62, (95% CI= -0.76 a -0.43)

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	componente modificado que permite o uso de membros superiores versus o TUG.					
TVC						
Fernandez-Huerta *et al*., (2019)	Nesse teste de velocidade de caminhada de 10 m (TVC10M), cada participante percorreu uma distância total de 14 m, composta por 2 m de aceleração, seguidos de 10 m de caminhada medidos com cronômetro e, por fim, 2 m de desaceleração. começam com um sinal verbal quando o participante está em pé e ambos os pés estão atrás da linha de partida. Cada participante foi instruído a caminhar o mais rápido possível, sem correr ou parar em direção à linha de chegada. Neste teste de velocidade de caminhada de 4 m (TVC4M), o percurso total foi 8 m, começou com 2 m de aceleração, seguido de 4 m de tempo de caminhada medido e depois 2 m de desaceleração. Cada participante foi instruído a caminhar o mais rápido possível, sem correr ou parar em direção à linha de chegada. O objetivo deste estudo foi avaliar a confiabilidade do teste de velocidade de caminhada de 10 m e 4 m cronometrados em	136 (M= 33 e H= 103)	72.83 (5.90)	Da comunidade	Velocidade (m/s)	[a]TVC4M Confiabilidade teste-reteste: ICC= 0.959 (95% CI=0.943-0.971) SEM= 0.067 MDC 95%= 0.185 [b]TVC10M Confiabilidade teste-reteste: ICC= 0.976 (95% CI=0.966-0.983) SEM= 0.053 MDC 95%= 0.146 [c]Confiabilidade da medição em diferentes distâncias (ICC entre as médias do teste de TVC4M e TVC10M): ICC= 0.867 (95% CI= -0.813-0.905) SEM= 0.082 MDC 95% = 0.227 Tempo do reteste: 1 minuto

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	comparação com os mesmos testes não cronometrados.					
Forte *et al*., (2021)	Nesse teste de velocidade de caminhada de 7 m foi utilizado portões de medição (Smart-speed, Fusion Sport, Coopers Plains, Austrália) colocados 3 m após a largada e 2 m antes da linha de chegada para permitir aceleração e evitar desaceleração antes do final do percurso. Os participantes caminharam um total de 12 m em (1) velocidade habitual, descrita aos sujeitos "como a velocidade com que caminhariam até as lojas" e em (2) velocidade máxima "o mais rápido possível sem correr". A partir daí, a velocidade máxima foi medida durante a caminhada nas seguintes condições: (a) em um caminho de largura reduzida de 25 e 15 cm, (b) pegando dois objetos (pequenos pesos manuais de 0,25 kg cada) colocados a 2 e 4 m do primeiro ponto de cronometragem e a 50 cm da linha média da pista; (3) ultrapassar duas barreiras (45 cm de largura e 15 e 45 cm de altura) colocadas sucessivamente na linha média da pista a 2 e 4 m do primeiro portão de cronometragem; (4) igual a (3), mas usando óculos	52 (M= 30 e H= 22)	69.7 (3.2)	Da comunidade	Velocidade (m/s)	Sessão 1 intra sessão (E1 vs E2): [a]VC habitual: ICC= 0.935 (CI 95%= 0.88-0.96) CV (CI 95%) = 3.6 (2.0–5.2) [d]VC máxima: ICC= 0.886 (CI 95%= 0.80-0.94) CV (CI 95%) = 3.5 (1.0–6.0) Sessão 2 intra sessão (E1 vs E2): [b]VC habitual: ICC= 0.949 (CI 95%= 0.91-0.97) CV (CI 95%) = 2.9 (1.6–4.1) [e]VC máxima: ICC= 0.921 (CI 95%= 0.86-0.96) CV (CI 95%) = 3.1 (1.5–4.3) S1 vs S2 intra sessão (média do E1 e E2): [c]VC habitual: ICC= 0.771 (95% CI= 0.632-0.862) MDC= 0.18 (12.9) [f]VC máxima: ICC= 0.628 (95% CI= 0.431-0.768) MDC= 0.27 (14.4) Em termos práticos, estes indicaram que eram necessárias alterações entre ≥0,18 m/s e 0,30 m/s para ter 95% de certeza de que a alteração não se devia a um erro de medição.

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	escuros; (5) carregar uma caixa de papelão vazia cobrindo a visão dos pés. O objetivo foi avaliar a confiabilidade absoluta intra e entre testes da velocidade de caminhada realizada em condições básicas e complexas.					Os gráficos de Bland e Altman mostraram as diferenças de velocidade de caminhada entre T1 e T2 plotados em relação as pontuações médias: LoA superior = 0.13 a 0.37 m/s LoA inferior= -29 a -0.49 m/s
Dewhurst *et al*., (2014)	Teste de velocidade máxima de caminhada de 6 m. Os participantes partiram de uma posição estática e caminharam o mais rápido que puderam até o final de um percurso de 9 m. Marcadores visíveis foram colocados no início e no 6 e 9 m. O tempo gasto desde o início até os 6 m foi registrado usando um cronômetro (Seiko, SO-52-4000, Tóquio, Japão) e a velocidade foi calculada. O objetivo do presente estudo foi avaliar a confiabilidade e sensibilidade intra sessão.	71 mulheres	71.7 (7.3)	Da comunidade	Capacidade funcional para tarefas da vida cotidiana	Confiabilidade teste-reteste intra dias (variáveis entre ensaios): 2 vs 1 ICC= 0.89 (95% CI= 0.82-0.93) 3 vs 2 ICC: 0.90 (95% CI= 0.85-0.94) Tempo do reteste: Todos os testes foram feitos em uma única sessão com 1 minuto de descanso entre as tentativas.
Adell *et al*., (2013)	Teste de velocidade máxima de caminhada de 10 m com fases de aceleração e desaceleração de 2 m cada. Os sujeitos foram instruídos a caminhar até a primeira linha e aumentar para velocidade máxima ao cruzar a primeira linha até cruzar a segunda linha. O avaliador caminhou ao lado do participante, iniciou a	31 (M= 25 e H= 6)	Média de 89 anos	Da comunidade	Desempenho físico	Confiabilidade teste-reteste (ensaio 1 vs ensaio 2): ICC= 0.86 Valor médio: t1: 0.97 m/s (SD= 0.30 m/s) t2: 0.95 m/s (SD= 0.29 m/s) Diferença média: -0.03 m/s (SD= 016 m/s) LoA= -0.33 a 0.27 m/s CV= 11.4%

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	cronometragem com um cronômetro digital quando o primeiro pé do sujeito cruzou a linha de partida e parou a cronometragem quando o primeiro pé cruzou a segunda linha. O objetivo deste estudo foi investigar a confiabilidade da velocidade máxima de caminhada de 10 m por meio do teste-reteste em idosos residentes em uma unidade residencial.					
Giné-Garriga *et al.*, (2010)	O teste TGUG modificado é uma ferramenta de avaliação da função física que mede o equilíbrio e a marcha com uma tarefa dupla: realizar uma tarefa cognitiva (contar regressivamente de 15 a 0) e uma tarefa física (andar em círculos) durante a caminhada é medido o tempo total necessário para realizar o teste (TT), o tempo decorrido entre o momento em que o sujeito chutou a bola e passou a linha dos 8 m (linha do kick-8 m) e o tempo desde o momento em que o sujeito cruzou a linha dos 8 m e retornou à cadeira (TT-kick 8 m). Os testes de comparação incluíram: velocidade de marcha rápida e normal (FGS e NGS), registrando o tempo que cada participante levou para caminhar	37 mulheres	72.3 (5.5)	Da comunidade	Desempenho físico correlacionado com declínio cognitivo (Marcha com dupla tarefa Cognitiva e motora)	Validade concorrente Componentes do TGUG modificado e medidas de desempenho físico: [a]Kick-8 m vs. NGS: r= 0.776 [b]Kick-8 m vs. FGS: r= 0.686 [c]TT-kick 8 m vs. NGS: r= 0.736 [d]TT-kick 8 m vs. FGS: r= 0.660

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	os 8 m centrais de um percurso de 12 m e dividindo a distância (8 m) pelo tempo da velocidade da marcha.					
Sayers *et al.*, (2006)	Teste de caminhada individualizada de 400 m (400 m W). Uma volta era de aproximadamente 100 m (-25 m de cada lado), portanto os 400 m consistiam em quatro voltas. Os participantes foram instruídos a caminhar em um ritmo que pudessem manter sem esforço excessivo até completarem os 400 m ou não conseguirem mais continuar. Para aqueles que não completaram o teste, foram registrados o horário em que o teste foi interrompido, o motivo da interrupção do teste e a distância percorrida. O objetivo do presente estudo foi avaliar a validade concorrente do teste recém-desenvolvido (400 m W) em comparação com o score do Short Physical Performance Battery (SPPB) e de seus componentes (levantar-se da cadeira, velocidade de caminhada de 4 m e score de equilíbrio). Um objetivo secundário deste estudo foi examinar o desempenho da caminhada em metros em indivíduos de alto funcionamento que tiveram bom	101 (M= 64 e H= 37).	80.8 (0.4)	Da comunidade	Mobilidade funcional	Validade concorrente (score SPPB e velocidade de caminhada 400 m W): [a]400 m W vs SPPB: r= 0.74 Correlações entre componentes do SPPB vs 400 m W: [b]400 m W vs Velocidade de caminhada de 4 m: r= 0.84 [c]400 m W vs tempo de levantar-se da cadeira: r= 0.53 67 participantes conseguiram completar os 400 m W (26 homens, 41 mulheres) e 34 não conseguiram (11 homens, 23 mulheres).

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	desempenho no SPPB (pontuação resumida 10 ou 12).					
Özden *et al.*, (2022)	Teste de caminhada para trás de 3 m (3MBWT). O início e o final do campo de 3 m foram marcados com fita colorida. Os pacientes foram solicitados a posicionar os calcanhares no nível da faixa horizontal da linha de partida. Os pacientes foram autorizados a olhar para trás. O teste foi realizado uma vez e o tempo foi registrado com cronômetro. Participantes foram instruídos a caminhar confortavelmente, com segurança, mas o mais rápido possível e parar na linha de chegada. O Teste de caminhada de 50 pés/15.24 m (50FWT) os participantes completam uma linha reta a aproximadamente 15,24 m (50 pés) de distância em velocidade normal de caminhada. O percurso foi determinado com fita plana e colorida medindo uma distância de 15,24 m. O tempo decorrido durante este teste é registrado. O objetivo do estudo foi avaliar a confiabilidade teste-reteste e validade concorrente do 50FWT com TUG e teste Five Times Sit to Stand (FTST) e 3MBWT com	65 (M= 34 e H= 31)	68.9 (3.7)	Da comunidade	Desempenho físico (equilíbrio estático e dinâmico, bem como força muscular adequada)	[a]3MBWT Confiabilidade teste-reteste: ICC= 0.940 (95% CI= 0.90–0.96) SEM= 0.55 MDC 95%= 1.52 Tempo do reteste: 1 hora Validade concorrente entre 3MBWT com TUG e FTST (Spear-man correlation coefcient (r)). [b]3MBWT vs TUG= 0.649** [c]3MBWT vs FTST= 0.238 Reteste [b]3MBWT vs TUG= 0.645** [c]3MBWT vs FTST= 0.217 [d]50FWT Confiabilidade teste-reteste: ICC= 0.820 (95% CI= 0.72–0.88) SEM= 1.28 MDC= 3.54 Validade Concorrente entre 50FWT com TUG e FTST (Spear-man correlation coefcient (r)). Teste [e]50FWT vs TUG= 0.550** [f]50FWT vs FTST= 0.215

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	TUG e teste Five Times Sit to Stand (FTST).					Reteste [e]50FWT vs TUG= 0.596** [f]50FWT vs FTST= 0.260*
Criss *et al.*, (2023)	Velocidade de caminhada lenta/walking at a slow speed (slowWS). Cada velocidade de caminhada (SSWS (velocidade de caminhada auto selecionada), MWS (velocidade máxima de caminhada) e velocidade lenta de caminhada (slowWS)) foi avaliada duas vezes em uma superfície plana de 20m, com áreas de aceleração e desaceleração de 5m e os 10m intermediários cronometrados com cronômetro. O objetivo deste estudo foi investigar a confiabilidade teste-reteste e entre avaliadores, erro padrão de medida (SEM), alteração mínima detectável no intervalo de confiança de 95% (MDC95) e correlação de slow WS com SSES e MWS.	110 (M= 87 e H= 23)	75.87 (7.17)	Da comunidade	Velocidade	Confiabilidade teste-reteste entre os ensaios SlowWS 1 e 2: [a]Avaliador 1: ICC= 0.971 (95% CI= 0.958-0.980) [a]Avaliador 2: ICC= 0.974 (95% CI= 0.962-0982) Confiabilidade entre avaliadores Ensaio 1 e 2: [b]Ensaio 1: ICC= 0.996 (95% CI= 0.994-0.997) [b]Ensaio 2: ICC= 0.997 (95% CI= 0.995-0.998) SEM= 0,014-0,015 m/seg MDCs= 0,04 m/seg. Tempo do reteste: Não relatado Validade concorrente (Correlações paramétricas de Pearson para variáveis de nível de intervalo/razão ou correlações não paramétricas de Spearman para variáveis de nível ordinais): [c]slowWS vs SSWS= .186 [d]slowWS vs MWS= 0.50 [f]MWS vs SSWS= .783* [e]slowWS vs 5 Times Sit to Stand Test (5xSTS) = .082 [g]MWS vs 5xSTS=. -691 [h]SSWS vs 5xSTS=. -564*

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
Saito *et al.*, (2022)	Teste de caminhada de 10 m com app de smartphone. Os próprios participantes do estudo mediram a velocidade de caminhada por meio de um smartphone equipado com o aplicativo de índice ME-BYO para pesquisa. O índice ME-BYO para uso geral é implementado no aplicativo My ME-BYO Record, que pode ser baixado na App Store. Os participantes seguiram as instruções na tela para caminhar 10 m em sua velocidade habitual e o tempo de caminhada foi medido. O local era uma área interna plana com zonas preliminares e de desaceleração de 4 m. Os participantes caminharam até o ponto final do caminho de desaceleração (18 m) sem parar ao final dos 10 m. Este estudo teve como objetivo verificar a validade e confiabilidade teste-reteste do teste de caminhada de 10 m medido por um app de smartphone medido pelos participantes vs. Instrutores fitness e vs. O TSLC 5 vezes.	40 (M= 20 e H= 20	74.9 (5.2)	Da comunidade	Velocidade (m/s)	[a]Confiabilidade teste-reteste: ICC= 0.712 (95% CI= 0.571-0.823) Tempo do reteste: Não relatado Validade de critério (Índice ME-BYO medida pelos participantes vs instrutores de fitness): [b]Segundo ensaio: r = 0.862 (95% CI = 0.753-0.925) Média dos três ensaios: r= 0.961 (95% CI= 0.927-0.979) [c]Índice ME-BYO (participantes) vs five-time sit-to-stand test: Segundo ensaio: r = −0.572 (95% CI = −0.750-0.317) Média dos três ensaios: r = −0.579 (95% CI = −0.754-0.326)

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
TPMI						
Balachandran *et al.*, (2021)	Teste de potência sentar levantar-se (STSp) uma cadeira com altura de 45 cm e um transdutor de posição linear foram usados para avaliar o pico de potência. Os participantes foram instruídos a começarem sentados com os braços cruzados sobre o peito e se levantar o mais rápido possível. O pico de potência (velocidade vertical [m/s] e massa movida [kg] para a parte do teste em pé) mais alto entre 3 sentar-se e levantar com 1 minuto de descanso entre as posições foi usado para análise. O objetivo do estudo é investigar a validade de construto, confiabilidade e erro de medição do STSp.	51 (M= 32 e H= 19)	71.3 (5.7)	Da comunidade	Potência de membros inferiores correlacionado com o teste de potência usando um leg press pneumático (LP) Desempenho correlacionado com o SPPB que inclui um teste de equilíbrio, velocidade habitual de caminhada e teste de levantar da cadeira; Teste Timed Up and Go (TUG) em ritmos normais e rápidos e medidas de resultados relatadas pelo paciente (PROMs), por meio do questionário de função física e mobilidade do Sistema de Informação de Medição de Resultados Relatados pelo	Confiabilidade teste-reteste: ICC= 0.96 (95% CI= 0.93-0.97) SEM= 70.4 W LoA= -187.9 - 201.0 W Tempo do reteste: Em duas ocasiões no período de 1 semana (mínimo) e 2 semanas (máximo) por 3 avaliadores em 36 participantes. Validade de construto: a. A potência de pico do STSp apresentou alta correlação de 0,90 (r Pearson) com a potência do leg press pneumático (LP); b. Para resultados de desempenho físico, o STSp mostrou correlações semelhantes ou superiores em comparação com o teste LP conforme hipotetizado: SPPB (r STSp= 0,41 vs. LP= 0,29); cadeira em pé (STSp= -0,44 vs. LP= -0,35); TUG normal (STSp= - 0,37 vs. LP= -0,29); TUG Rápido (STSp= -0,41 vs. - LP= 0,34), e equilíbrio (0,33 vs. 0,22). Para PROMs, em contraste com a hipótese, LP apresentou maior correlação com questionários de mobilidade (0,34 vs. 0,38) e

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
					Paciente (PROMIS), desenvolvido pelo National Institutes of Hea.lth (NIH)	questionário de função física (0,41 vs. 0,48); c. Análises exploratórias mostraram correlações de Spearman e pico de potência em relação ao peso corporal (W/kg) para ser consistente com nossa análise primária. Para a validade discriminante, como esperado, os homens apresentaram maior potência de pico STSp em comparação com as mulheres (A = 492 W, p <0,001, d de Cohen= 2.0.
Sherwood *et al.*, (2020)	Sit-to-Stand (STS) de 5 movimentos máximos com sistema de transdutor de posição linear foram registrados. os participantes foram instruídos a sentar-se em uma cadeira com os pés afastados na largura dos ombros, os joelhos posicionados diretamente sobre os pés e os braços cruzados sobre o peito com as mãos nos ombros opostos, até que o examinador os instruísse a se levantarem. uma posição sentada. A mesma altura da cadeira com encosto de madeira (0,41 m) foi utilizada para todos os testes. As instruções para a tarefa STS foram padronizadas e incluíam o seguinte: “Levante-se o mais	48 (M= 36 e H= 12)	77.6 (11.1)	Da comunidade	Potência de membros inferiores em comparação com a análise da videografia 2D Dartfish (DF)	Concordância absoluta Dartfish (DF) entre avaliadores: [a]Velocidade DF/avaliador 1 vs DF/avaliador 2: ICC = 0.95 (95% CI= 0.92-0.97) [b]Potência DF/avaliador 1 vs DF/avaliador 2: ICC = 0.95 (95% CI= 0.92-0.97). Validade concorrente (Dartfish vs GymAaware): [c]Velocidade (concordância absoluta): ICC = 0.94 (95% CI= 0.89-0.97) [c]Velocidade (consistência): ICC = 0.95 (95% CI= 0.91-0.97)

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	rápido possível". Os participantes realizaram uma a duas repetições práticas antes do início, após as quais um vídeo foi capturado. Os objetivos do estudo foram analisar a confiabilidade e validade das medidas de velocidade e potência do transdutor de posição linear GymAware durante o sentar levantar-se, em comparação com a análise da videografia 2D Dartfish (DF).					[d]Potência (concordância absoluta): ICC = 0.98 (95% CI= 0.96-0.99) [d]Potência (consistência): ICC = 0.98 (95% CI= 0.96-0.99)
Cruvinel-Cabral *et al.*, (2018)	My jump app. Todos os participantes realizaram quatro saltos contramovimento (CMJ) na sessão de familiarização. Após aquecimento padrão, cada participante realizou três CMJs máximos em um tapete de contato personalizado conectado a um computador com software específico (Chronojump, versão 1.6.2; Boscosystem, Barcelona, Espanha). Este equipamento foi anteriormente demonstrado como válido e confiável (Pagaduan & De Blas, 2004; De Blas et al., 2012). Paralelamente, o salto foi registrado pelo mesmo pesquisador com um celular (iPhone 7; Apple. Cupertino, CA, EUA) a uma taxa de amostragem de 240 Hz, por meio do aplicativo My Jump. O	41(M= 29 e H= 12)	H= 73,2 (6,4) e M= 69,4 (8,9)	13 da instituição e 28 da comunidade	Força muscular explosiva ou potência de membros inferiores	Confiabilidade teste-reteste entre avaliadores: Confiabilidade relativa ICC= 0.948 (95% CI= 0.913–0.970) Confiabilidade absoluta Erro típico de medição (TEM)= 1.150 Coeficiente de variação (CV%) = 10.096 As diferenças entre os dois métodos utilizando o salto mais alto foi de 0,096 cm com limites de concordância de (LoA) de 0,3255 cm a -0,5177 cm. Tempo do reteste: 3 saltos sucessivos Validade de concorrente (My Jump vs Tapete de Contado Chronojump). Salto mais alto,

Tabela 1 - Cont.

Estudo	Descrição do teste e afins	População			Construto medido	Propriedades dos instrumentos de medição de resultados
		n	idade	tipo		
	objetivo do estudo foi verificar a validade e confiabilidade do app "My Jump" para avaliação da altura do salto com contramovimento (SCM/CMJ) em uma população idosa.					média dos 3 saltos e todos os saltos: r = 0.999; P = 0.000
Giné-Garriga *et al.*, (2010)	O teste TGUG modificado é uma ferramenta de avaliação da função física que mede o equilíbrio e a marcha com uma tarefa dupla (já descrito acima) Os testes de comparação incluíram: (a) cinco elevações de cadeira (Cr) com uma cadeira de 42 cm de altura do assento ao chão com os braços do sujeito cruzados sobre o peito; e (b) duas forças de contração isométricas máximas de quadríceps e isquiotibiais (CVM) para cada perna realizadas durante 5 s com um período de descanso de 45 s entre cada uma, obtidas em um ângulo de articulação do joelho de 60 (0 = extensão total do joelho) usando um dinamômetro isocinético (ConTrex.Human Kinetics 1.7.1, Hans e Rüth), e utilizando o maior valor para análise dos dados.	37 mulheres	72.3 (5.5)	Da comunidade	Desempenho físico correlacionado com declínio cognitivo (Força das extremidades inferiores do corpo)	Componentes do TGUG modificado e medidas de desempenho físico: [a]SU vs. Cr: r= 0.691 [b]BT vs. Cr: r= 0.659 [c]Right knee extensor, MVCc vs. SU: r= 0.692 [d]Left knee extensor MVCc vs. SU: r= 0.597 [e]Right knee extensor MVCc vs. BT: r= 0.640 [f]Left knee extensor MVCc vs. BT: r= 0.507

Nota: TLSC= teste de levantar e sentar da cadeira, TVC= teste de velocidade de caminhada, TPMI= teste de potência de membros inferiores, ICC= intraclass correlation coefficient, CI= confidence interval, SEM = standard error of measurement, CV= coefficient of variation, MDC= minimal detectable change, SRD= smallest real difference, MDD= minimal detectable difference, LoA= limits of agrément. Fonte: Elaborado pelo autor.

4.3 Avaliação da qualidade metodológica (*COSMIN Risk of Bias Checklist*)

4.3.1 *TUG* - *COSMIN Risk of Bias Checklist*

A análise abrangente dos estudos relacionados ao teste *TUG* revelou uma investigação aprofundada em suas propriedades de medição. Dos oito estudos examinados, seis concentram-se na avaliação da confiabilidade e erro de medição (GALHARDAS; RAIMUNDO; MARMELEIRA, 2020; NEPAL; BASAULA; SHARMA, 2020; COLLADO-MATEO *et al*., 2019; LE BERRE *et al*., 2016; LEE *et al*., 2016; DEWHURST; BAMPOURAS, 2014). Notavelmente, apenas o estudo de Le Berre *et al*., 2016, aborda não apenas a confiabilidade e erro de medição, mas também a validade de construto. Adicionalmente, dois estudos direcionam sua atenção exclusivamente à validade de critério (CHAN *et al*., 2016; GINÉ-GARRIGA *et al*., 2010).

Ao explorar a confiabilidade, Le Berre *et al*., (2016) e Nepal, Basaula e Sharma (2020), obtiveram avaliações de muito bom (V), enquanto Galhardas, Raimundo e Marmeleira (2020) e Lee *et al*., (2016), foram avaliados como adequados (A) e Collado-Mateo *et al*., (2019) e Dewhurst e Bampouras (2014), receberam classificação duvidosa (D). Na análise do erro de medição, somente Le Berre *et al*., (2016), recebeu uma avaliação muito boa (V), enquanto Galhardas, Raimundo e Marmeleira (2020) e Lee *et al*., (2016), foram classificados como adequados (A) e (COLLADO-MATEO *et al*., 2019; DEWHURST; BAMPOURAS, 2014; NEPAL; BASAULA; SHARMA, 2020), foram considerados duvidosos (D).

No que diz respeito à validade, (CHAN *et al*., 2016; GINÉ-GARRIGA *et al*., 2010; LE BERRE *et al*., 2016), alcançaram avaliações muito boas (V) para validade de critério e construto, respectivamente. Todos esses resultados detalhados das propriedades de medição do *TUG* estão meticulosamente documentados na Tabela 2, proporcionando uma visão abrangente e crítica dessas avaliações.

4.3.2 TSLC - *COSMIN Risk of Bias Checklist*

Na análise abrangente das propriedades de medição do TSLC, conduzida por meio de três estudos distintos, emergiram resultados discerníveis. Dois desses estudos (COLLADO-MATEO *et al*., 2019; LE BERRE *et al*., 2016), concentraram-se na avaliação da confiabilidade e erro de medição, indicando variações pertinentes em seus resultados. Enquanto o estudo de Collado-Mateo *et al*., (2019), revelou uma confiabilidade duvidosa (D), o estudo de Le Berre e*t al*., (2016), destacou-se com uma classificação muito boa (V). No que tange à avaliação do erro de medição, Colado-Mateo *et al*., (2019), obteve uma classificação duvidosa (D), enquanto Le Berre *et al*., (2016), apresentou uma avaliação muito boa (V). O terceiro estudo, conduzido

por Chan *et al*., (2016), concentrou-se na validade de critério, obtendo uma avaliação muito boa (V). Todos os resultados pertinentes às propriedades de medição do TSLC estão minuciosamente apresentados na Tabela 2, proporcionando uma visão holística dessas análises críticas.

4.3.3 TVC - *COSMIN Risk of Bias Checklist*

A investigação das propriedades de medição do TVC envolve nove estudos, dos quais sete abordam a confiabilidade e o erro de medição (ADELL *et al*., 2013; CRISS *et al*., 2023; DEWHURST *et al*., 2014; FERNÁNDEZ-HUERTA *et al*., 2019; FORTE *et al*., 2021; ÖZDEN *et al*., 2022; SAITO *et al*., 2022), enquanto os três restantes concentram-se na validade de critério (GINÉ-GARRIGA *et al*., 2010; SAYERS *et al*., 2006) e Saito *et al*., (2022), que, além da validade de critério, avaliou também a confiabilidade.

Ao explorar a confiabilidade, três estudos (ADELL *et al*., 2013; FORTE *et al*., 2021; ÖZDEN *et al*., 2022) receberam avaliação de adequado (A), enquanto quatro estudos (CRISS *et al*., 2023; DEWHURST *et al*., 2014; FERNÁNDEZ-HUERTA *et al*., 2019; SAITO *et al*., 2022) foram avaliados como duvidosos (D). Na análise do erro de medição, três estudos (ADELL *et al*., 2013; FORTE *et al*., 2021; ÖZDEN *et al*., 2022) foram classificados como adequados (A), enquanto três estudos (CRISS *et al*., 2023; DEWHURST *et al*., 2014; FERNÁNDEZ-HUERTA *et al*., 2019) receberam classificação duvidosa (D).

Quanto à validade de critério, os estudos de Giné-GArriga *et al*., (2010) e Sayers *et al*., (2006), foram avaliados como muito bons (V), contrastando com o estudo de Saito *et al*., (2022), que recebeu avaliação duvidosa (D). Detalhes completos de todas as avaliações das propriedades de medição do TVC estão registrados de maneira abrangente na Tabela 2, proporcionando uma visão compreensiva dessas análises críticas.

4.3.4 TPMI - *COSMIN Risk of Bias Checklist*

A avaliação das propriedades de medição do TPMI envolve quatro estudos distintos. Desses, dois estudos (BALACHANDRAN *et al*., 2021; CRUVINEL-CABRAL *et al*., 2018) dedicaram-se à análise da confiabilidade e erro de medição, enquanto três estudos (CRUVINEL-CABRAL *et al*., 2018; GINÉ-GARRIGA *et al*., 2010; SHERWOOD *et al*., 2020) focaram na validade de critério, sendo que o último estudo, adicionalmente, avaliou a confiabilidade. Além disso, um estudo (BALACHANDRAN *et al*., 2021) concentrou-se e no teste de hipóteses pelos métodos de validade convergente e discriminativa para a confirmação da propriedade de medição validade de construto.

Na avaliação da confiabilidade, os resultados variaram, com (BALACHANDRAN *et al*., 2021) sendo avaliado como adequado (A), (SHERWOOD *et al*., 2020) com muito bom (V), e (CRUVINEL-CABRAL *et al*., 2018) sendo classificado como inadequado (I). No que diz respeito ao erro de medição, (CRUVINEL-CABRAL *et al*., 2018) foi avaliado como inadequado (I), enquanto (BALACHANDRAN *et al*., 2021) obteve avaliações de adequado (A) e muito bom (V), especificamente no teste de hipóteses para a validade de construto. Na avaliação da validade de critério, todos os três estudos (CRUVINEL-CABRAL *et al*., 2018; GINÉ-GARRIGA *et al*., 2010; SHERWOOD *et al*., 2020) receberam classificações consistentes de muito bom (V).

Os resultados de avaliação das propriedades de medição do TPMI são detalhadamente apresentados na Tabela 2, essa compilação oferece uma visão holística e crítica dessas análises, contribuindo para a compreensão rigorosa do desempenho do teste em diferentes contextos de pesquisa.

Tabela 2 - Avaliação da qualidade metodológica dos estudos: *COSMIN Risk of Bias Checklist*

Testes	Confiabilidade				Erro de Medição				Validade de Critério				Teste de Hipóteses para Validade de Construto			
	V	A	D	I	V	A	D	I	V	A	D	I	V	A	D	I
TUG																
Collado-Mateo *et al.*, (2019)			x				x									
Galhardas *et al*., (2020)		x				x										
Nepal *et al*., (2020)	x						x									
Chan *et al.*, (2016)									x							
Le Berre *et al.*, (2016)	x				x								x			
Lee *et al*., (2016)		x				x										
Dewhurst *et al*., (2014)			x				x									
Giné-Garriga *et al.*, (2010)									x							
TLSC																
Collado-Mateo *et al.*, (2019)			x				x									
Chan *et al.*, (2016)									x							
Le Berre *et al.*, (2016)	x				x								x			

Tabela 2 - Cont.

Testes	Confiabilidade				Erro de Medição				Validade de Critério				Teste de Hipóteses para Validade de Construto			
	V	A	D	I	V	A	D	I	V	A	D	I	V	A	D	I
TVC																
Fernandez-Huerta et al., (2019)			x				x									
Forte *et al.*, (2021)		x				x										
Dewhurst *et al.*, (2014)			x				x									
Adell *et al.*, (2013)		x				x										
Giné-Garriga *et al.*, (2010)									x							
Sayers *et al.*, (2006)									x							
Özden *et al.*, (2022)		x				x			x							
Criss *et al.*, (2023)			x				x		x							
Saito *et al.*, (2022)			x						x							
TPMI																
Balachandran *et al.*, (2021)		x				x							x			
Sherwood *et al.*, (2020)	x								x							
Cruvinel-Cabral *et al.*, (2018)				x				x	x							
Giné-Garriga *et al.*, (2010)									x							

Nota: V= *very good*; A= *adequate*; D= *doubtful*; I= *inadequate*. Fonte: elaborado pelo autor.

4.4 Critérios de avaliação para as boas propriedades de medição

Nesta etapa, há alguns pontos importantes que precisam ser esclarecidos antes da leitura dos resultados. (i) Há estudos que avaliaram mais de um teste, alguns avaliaram o mesmo teste com alguma modificação, e outros, são testes completamente diferentes, há também testes que realizaram mais de uma medida para uma propriedade de medição de algum domínio (p. ex. para a confiabilidade, alguns estudos realizaram teste-reteste para avaliar a confiabilidade intra dia e entre dias ou intra avaliador e entre avaliadores e testes que foram comparados com mais de um teste para avaliar as propriedades de medição tanto da validade de critério como da validade de construto). Para estas ocasiões, consideramos testes e/ou medidas de resultados diferentes que deveriam ser avaliados individualmente como os demais.

(ii) Alguns estudos realizaram teste-reteste com mais de dois ensaios para avaliar a concordância de uma determinada medida, no entanto, para nossas análises, consideramos sempre o ensaio um e ensaio dois, embora alguns estudos apresentem mais ensaios. Tendo em vista todos esses expostos supracitados, as tabelas 1 e 3 têm seus testes com suas medidas de resultados e referências respectivamente identificados com letras sobrescritas (p. ex. [a, b, c, d, e...]) para uma melhor identificação e entendimento.

(iii) Na avaliação das propriedades de medição erro de medição, o manual do usuário *COSMIN methodology for systematic reviews* considera limites de *LoA*, *SDC* e *MIC* como medidas adequadas para avaliar o domínio erro de medição (PRINSEN, *et al.*, 2018; PRINSEN *et al.*, 2016; TERWEE *et al.*, 2007). No entanto, todos os estudos avaliados para esta propriedade de medição, não realizaram a *MIC*, os próprios autores da iniciativa *COSMIN* publicaram um artigo específico sobre a explicação do que realmente seria o *MIC*.

Em conformidade com Terwee *et al.*, (2021), o *MIC* (alteração mínima importante) é definido como um ponto crítico para detectar variações mínimas significativas na experiência do paciente ao longo do tempo. Considerando que cada paciente possui uma percepção individual do que constitui uma mudança mínima relevante, o *MIC* pode ser entendido como a média desses pontos de referência individuais. Essa concepção de *MIC* envolve três elementos cruciais: em primeiro lugar, estabelece um limite para mudanças mínimas que os pacientes percebem como significativas (seja de melhora ou deterioração). Em segundo lugar, delimita uma alteração considerada importante na perspectiva do paciente. E, por último, refere-se a mudanças ocorridas dentro do paciente ao longo do tempo.

Sendo assim, consideramos que todos os estudos realizaram a menor alteração detectável (*SDC*) apresentadas também com terminologias diferentes como alteração mínima detectável (*MDC*), menor diferença real (*SRD*) e diferença mínima detectável (*MDD*), no entanto, essas não são o *MIC*. O *SDC*, *MDC*, *SDR* e *MDD* diz respeito a menor alteração na pontuação que pode ser detectada estatisticamente com algum grau de certeza (p. ex. 95 ou 90%), com base no *SEM* ou nos *LoA* de medidas de confiabilidade teste-reteste. E, portanto, como nenhum estudo definiu o *MIC*, todos os testes que avaliaram as propriedades de medição erro de medição, foram avaliados com indeterminado (?).

(iv) A taxonomia para propriedades de medição desenvolvida pela iniciativa *COSMIN*, subdivide o domínio validade em três propriedades de medição, dentre elas, as que nos interessa, validade de critério e validade de construto. Para a validade de critério, não há nenhum aspecto desta propriedade de medição descrita pelo *COSMIN*, no entanto, alguns autores subdividem a validade de critério em duas: validade preditiva quando o critério se situa

no futuro, e validade concorrente quando é contemporâneo. Ou seja, se um teste é aplicado e seus resultados são comparados com um critério aplicado um tempo depois, obtém-se a validade preditiva, e se ambos os testes são aplicados ao mesmo tempo, tem-se a validade concorrente. Sendo assim, todos os estudos incluídos em nossa revisão para a propriedade de medição validade de critério, realizaram a validade concorrente (POLIT, 2015).

Para a validade de construto, há três aspectos desta propriedade de medição, dentre eles, o que nos interessa, o teste de hipóteses. A iniciativa *COSMIN* descreve dois métodos para confirmação da validade de construto pelo teste de hipóteses: validade discriminativa ou de grupos conhecidos e a validade convergente que foi o método utilizado por todos os estudos incluídos na avaliação para validade de construto, com exceção do estudo de Balachandran *et al*., (2021), que realizou ambos.

A validade convergente consiste na ausência de um instrumento 'padrão-ouro', a possibilidade de testar a validade convergente por meio da correlação das pontuações do instrumento focal com os escores de outro instrumento que avalie um construto similar. Assim, é possível verificar se o instrumento avaliado está fortemente correlacionado a outras medidas já existentes e válidas. Já a validade discriminativa, por sua vez, verifica se a medida em análise não possui relação inadequada com construtos distintos, ou seja, se está devidamente desvinculada de variáveis com as quais deveria se diferenciar (POLIT, 2015; MOKKINK *et al.*, 2010b).

(v) Alguns estudos realizaram correlação entre testes, mas não relataram qualquer intenção de avaliar a validade concorrente ou divergente, para estes casos, consideramos medidas de resultados de confiabilidade avaliada pelo teste paralelo ou formas equivalentes, que se refere à concordância entre dois ou mais instrumentos que medem o mesmo atributo, cuja aplicação ocorreu em tempos distintos, em intervalo curto de tempo e aplicados aos mesmos indivíduos (POLIT, 2015).

4.4.1 *TUG* - Critérios de avaliação para as boas propriedades de medição

Na análise abrangente do *TUG*, uma série de medidas de resultados foram consideradas para avaliação da confiabilidade, erro de medição, validade de critério e teste de hipóteses para validade de construto. Ao examinar 13 medidas para confiabilidade e 12 para erro de medição, das 13, somente uma foi avaliada como indeterminada (?) e as 12 restantes foram classificadas como suficientes (+), enquanto as 12 do erro de medição foram avaliadas como indeterminadas (?). Para a validade de critério, três medidas foram avaliadas, todas consideradas suficientes

(+), e no teste de hipóteses para validade de construto, uma medida foi classificada como indeterminada (?).

O estudo de Collado-Mateo *et al*., (2019), destacaram a confiabilidade teste-reteste entre cinco ensaios em uma única sessão do *TUG* medido por um cronômetro manual e um cronômetro portátil no qual apresentou ligeiramente uma melhor confiabilidade. Para os ensaios, sugere-se que o primeiro funciona como familiarização, evidenciado pelo tempo significativamente maior em comparação com o segundo. A quinta repetição, provavelmente influenciada pelo tempo de descanso de 1 minuto, mostra aumento significativo. Recomenda-se realizar no mínimo duas e no máximo quatro repetições em pessoas idosas, destacando que a terceira repetição é a mais consistente e pode ser apropriada para avaliações do *TUG*. Por fim, o estudo avaliou a correlação entre os diferentes métodos de cronometragem do *TUG* (manual e portátil) em pessoas idosas saudáveis da comunidade. Com três medidas de resultados, as propriedades de medição foram avaliadas separadamente.

Da mesma forma, os estudos de Galhardas, Raimundo e Marmeleira (2020) e Nepal, Basaula e Sharma (2020), centraram-se na confiabilidade teste-reteste, de dois ensaios realizados em duas sessões do tempo (s) de execução do *TUG* original em pessoas idosas institucionalizadas com mais de 75 anos e confiabilidade teste-reteste entre avaliadores em dois ensaios na mesma sessão de teste do *TUG* medido por estudantes de fisioterapia e *TUG* medido por cuidadores de pessoas idosas da comunidade respectivamente. Avaliando duas propriedades de medição cada, enquanto o estudo de Chan *et al*., (2016), focaram na validade concorrente entre o *TUG* medido por um aplicativo de smartphone e um sensor de força considerado uma condição de referência laboratorial em pessoas idosas da comunidade, avaliando uma propriedade de medição.

Ao explorar a confiabilidade Le Berre *et al*., (2016), realizou teste-reteste do *TUG* original e a validade convergente com o TSLC de 30 segundos modificado que permite o uso dos membros superiores em pessoas idosas longevas (82 - 98 anos) institucionalizados, avaliando três propriedades de medição. Da mesma forma, o estudo de Lee *et al*., (2016), examinou a confiabilidade teste-reteste do *TUG* medido por um cronômetro portátil e *TUG* com cronometragem baseada em carga nas distâncias de 3, 6 e 9 metros, porém, em pessoas idosas da comunidade, considerando seis medidas de resultados e suas respectivas propriedades de medição. Já o estudo de Dewhurst e Bampouras (2014), enfocou a confiabilidade intra dia com três ensaios de uma sessão de teste do *TUG* original (*Timed 8-Foot Up-and-Go*) de 2,44 m só em mulheres idosas da comunidade fisicamente ativas, avaliando duas propriedades de medição.

Ao investigar a validade concorrente Giné-Garriga *et al*., (2010), examinaram o *Timed Get Up and Go test* (*TGUG*) modificado com dupla tarefa (cognitiva e motora) e suas próprias fases e com outros testes, optou-se por separá-los em cada teste correspondente. Para o *TUG*, foi incluído a validade concorrente entre o tempo total para realizar o *TGUG* descrito como (*TT*) e os TVC normal e rápida. Todos esses resultados detalhados sobre as propriedades de medição do *TUG* estão disponíveis na tabela 3, oferecendo uma visão abrangente e crítica dessas análises.

4.4.2 TSLC - Critérios de avaliação para as boas propriedades de medição

A avaliação do TSLC envolveu a consideração de diversas medidas de resultados, abrangendo confiabilidade, erro de medição, validade de critério e teste de hipótese para validade de construto. Neste contexto, duas medidas de confiabilidade foram classificadas como suficientes (+), enquanto duas medidas para erro de medição foram avaliadas como indeterminadas (?). A validade de critério foi avaliada como suficiente (+), e o teste de hipótese para validade de construto foi classificado como indeterminado (?).

No estudo conduzido por Collado-Mateo *et al*., (2019), a confiabilidade teste-reteste do TSLC 30 segundos medidos automaticamente foi examinada em uma sessão com pessoas idosas saudáveis da comunidade, resultando em duas propriedades de medição avaliadas pelo mesmo teste. O estudo de Chan *et al*., (2016), concentrou-se na validade concorrente do TSLC 5 vezes medido por um aplicativo de smartphone, comparando-o com um sensor de força considerado referência laboratorial em pessoas idosas da comunidade, resultando em uma propriedade de medição avaliada pelo mesmo teste. Le Berre *et al*., (2016), investigaram a confiabilidade teste-reteste e a validade convergente entre o TSLC modificado com o uso dos membros superiores e o *TUG* em pessoas idosas longevas institucionalizadas, avaliando três propriedades de medição pelo mesmo teste. Detalhes abrangentes dessas análises e resultados estão disponíveis na tabela 3.

4.4.3 TVC - Critérios de avaliação para as boas propriedades de medição

A avaliação do TVC envolveu uma análise abrangente de 16 medidas para confiabilidade (15 consideradas suficientes [+] e 1 insuficiente [-]), 10 para erro de medição, 19 para validade de critério (6 consideradas suficientes [+] e 13 insuficientes [-]). A extensa revisão incluiu uma variedade de estudos que exploraram diferentes facetas do TVC, fornecendo uma compreensão aprofundada das suas propriedades de medição. No estudo de Fernandez-Huerta *et al*., (2019), instaurou três medidas de confiabilidade, sendo teste-reteste do TVC máxima de 4 e 10 metros, além da correlação entre as médias dos testes TVC4M e

TVC10M com fases cronometradas e fases não cronometradas (2 m de aceleração e 2 m de desaceleração) em pessoas idosas da comunidade. Embora a confiabilidade de ambos os testes de velocidade de caminhada seja excelente, o TVC4M apresenta uma concordância na análise de Bland-Altman ligeiramente inferior em comparação com o TVC10M, apresentando um resultado médio superior ao *MDC*.

Mais recentemente, Forte *et al*., (2021), focaram na confiabilidade intra e entre sessões, explorando diferentes velocidades de execução (habitual e rápida) do TVC de 7 m com fases não cronometradas de 3 m no início e 2 m no final em pessoas idosas da comunidade, resultando em seis medidas de resultados. Neste tocante, Dewhurst e Bampouras (2014), contribuíram com uma análise da confiabilidade intra sessão do TVC máxima de 6 m com fase não cronometrada de desaceleração de 3 m em mulheres idosas saudáveis da comunidade, destacando a importância da repetibilidade de apenas dois ensaios com a dispensa do ensaio de familiarização pela alta confiabilidade e sensibilidade entre os ensaios realizados, gerando uma medida de resultado avaliada. Já Adell, Wehmhörner e Rydwik (2013), contribuíram examinando a confiabilidade teste-reteste do TVC máxima de 10 m com fases não cronometradas de 2 m no início e no final em pessoas idosas da comunidade. Embora tenham apresentado valiosas informações sobre a confiabilidade, vale ressaltar que o estudo reportou apenas o resultado do *ICC* sem *CI*.

Há uma análise bem peculiar para o estudo de Giné-Garriga *et al*., (2010), expandiram o escopo, avaliando a validade concorrente entre o *TGUG* modicado com dupla tarefa (cognitiva e motora) e suas próprias fases e com outros testes, e optou-se por separá-las em cada teste correspondente. Em suas fases, há duas que correspondem ao TVC, um com tarefa cognitiva chamado de *Kick-8m* (tempo decorrido entre o momento em que o sujeito chutou a bola e ultrapassou a linha de 8 m e o TVC com tarefa motora chamado *TT-Kick-8m* (o momento em que o sujeito cruzou a linha de 8 m e retornou à cadeira).

Ambas as fases avaliaram a validade concorrente com o TVC normal e máxima de 8 m em um percurso total de 12 m, oferecendo uma compreensão mais detalhada das propriedades de medição do TVC. Sayers *et al*., (2006), trouxeram uma perspectiva adicional, examinando a validade concorrente entre o TVC individualizada de 400 m (*400 m W*) e os escores gerais do *Short Physical Performance Battery* (*SPPB*), além de seus componentes, como TSLC e TVC de 4 m. Essa abordagem multifacetada contribuiu para a compreensão abrangente de seis medidas avaliadas.

Os estudos de Özden *et al*., (2022), Criss *et al*., (2023) e Saito *et al*., (2022), forneceram uma visão holística, explorando diferentes aspectos da confiabilidade teste-reteste, validade

concorrente e comparações com outros testes, por meio de métodos variados, esses estudos enriqueceram a compreensão das propriedades de medição do TVC em diversas condições e contextos, como o uso de aplicativos de smartphone aplicados somente em pessoas idosas residentes da comunidade totalizando 6, 8 e 3 medidas de resultados para cada estudo respectivamente.

Neste contexto, Özden *et al*., (2022), avaliaram a confiabilidade teste-reteste de dois ensaios em uma única sessão em uma TVC máxima para trás de 3 m (*3MBWT*) e TVC habitual de 50 pés/15.24 m (*50FWT*) sem fases não cronometradas além de avaliar a validade concorrente dos testes *3MBWT* e *50FWT* comparando cada um com o *TUG* e TSLC 5 vezes. De forma parecida, Criss *et al*., (2023), avaliou a confiabilidade teste-reteste entre os ensaios e entre avaliadores com quantidade de sessão e tempo de reteste não relatados para o TVC lenta de 10 m (*slowWS*), com fases não cronometradas de 5 m no início e no final além de avaliar a validade concorrente entre *slowWS* com TVC auto selecionada de 10 m (*SSWS*), TVC máxima de 10 m (*MWS*) e TSLC 5 vezes, *MWS* com *SSWS* e TLSC 5 vezes e por fim *SSWS* com TSLC 5 vezes.

Adicionalmente, Saito *et al*., (2022), examinaram a confiabilidade teste-reteste com a quantidade de sessão e tempo de reteste não relatados para o TVC habitual de 10 m com fases não cronometradas de 4 m no início e final medido por um aplicativo de smartphone chamado de índice *ME-BYO* para o sistema *IOS*, além da validade concorrente do TVC de 10 m medido pelos participantes com o TVC de 10 m medido pelos instrutores fitness e TSLC 5 vezes com o uso do aplicativo. Os resultados detalhados dessas análises, apresentados na tabela 3, oferecem uma visão abrangente das boas propriedades de medição do TVC, destacando a complexidade e a heterogeneidade das avaliações realizadas.

4.4.4 TPMI - Critérios de avaliação para as boas propriedades de medição

A análise detalhada do TPMI revela uma abordagem abrangente, com ênfase em quatro medidas para confiabilidade avaliadas com suficientes (+), duas para erro de medição consideradas indeterminadas (?), nove para validade de critério (três consideradas suficientes [+] e seis insuficientes [-]), e um estudo realizou dez conjuntos de medidas, para avaliar a validade de construto pelo teste de hipóteses, sendo oito pelo método de validade convergente, uma para confiabilidade e uma para validade discriminativa, este conjunto de medidas resultou em uma avaliação geral do teste de hipóteses para validade de construto suficiente (+).

O estudo de Balachandran *et al*., (2021), destacou a confiabilidade teste-reteste e correlação entre os testes de potência sentar-levantar 3 vezes (*STSp)* medido por um transdutor

de posição linear (*LPT*) e *leg press* (*LP*) pneumático e, simultaneamente, explorou a validade convergente para a confirmação da validade de construto pelo teste de hipóteses através de medidas de desempenho físico (*SPPB*, TVC4M, teste de equilíbrio e *TUG*), medidas autorrelatadas (questionário de mobilidade e função) e para avaliar a validade discriminante, comparou o pico de potência do *STSp* entre mulheres e homens pessoas idosas da comunidade, confirmando hipóteses estabelecidas previamente, totalizando dez medidas de resultados para avaliar de forma geral a validade de construto.

Na mesma direção, o estudo de Sherwood *et al*., (2020), utilizando também o TSLC 5 vezes medido pelo sistema GymAware (*GA*) de *LPT* e por análise da videografia 2D *Dartfish* (*DF*), avaliou a confiabilidade entre avaliadores realizada com as medidas do *DF* para velocidade e potência, além de avaliar a validade concorrente para velocidade e potência entre *DF* e *GA* em pessoas idosas da comunidade. Este estudo, com quatro medidas de resultados avaliadas, enriqueceu a revisão com insights sobre as propriedades de medição do TPMI.

Avaliando a confiabilidade Cruvinel-Cabral *et al.*, (2018), explorou o teste-reteste entre avaliadores de 3 saltos contramovimento (*CMJs*) sucessivos em um tapete de contato e, paralelamente, o salto foi registrado pelo mesmo pesquisador com um celular (*iPhone 7; Apple. Cupertino, CA, EUA*) a uma taxa de amostragem de 240 Hz, por meio do aplicativo *My Jump*, e em seguida as medidas de ambos os equipamentos foram avaliadas para validade concorrente, resultando em três propriedades de medição avaliadas pelo mesmo teste.

Finalmente, Giné-Garriga *et al*., (2010), ofereceram uma perspectiva única, explorando a validade concorrente entre o tempo necessário para se levantar de uma cadeira sem usar os braços e chutar a bola (*SU*) e o tempo decorrido entre o momento em que o teste o sujeito chutou a bola até quando a bola ultrapassou a linha dos 8 m (*BT*), ambos considerados como construtos de força dos membros inferiores. Este estudo, com seis medidas de resultados avaliadas, adicionou nuances importantes às propriedades de medição do TPMI, completando a revisão com uma visão abrangente das boas práticas nessa área. Detalhes dessas análises e resultados estão disponíveis na tabela 3.

Tabela 3 - Critérios de avaliação por teste dos estudos para as boas propriedades de medição

Testes	Confiabilidade			Erro de Medição			Validade de Critério			Teste de Hipóteses para Validade de Construto		
	+	?	-	+	?	-	+	?	-	+	?	-
TUG												
Collado-Mateo *et al.*, (2019)[a]	x				x							
Collado-Mateo *et al.*, (2019)[b]	x				x							
Collado-Mateo *et al.*, (2019)[c]			x									
Galhardas *et al.*, (2020)	x				x							
Nepal *et al.*, (2020)	x				x							
Chan *et al.*, (2016)							x					
Le Berre *et al.*, (2016)	x				x						x	
Lee *et al.*, (2016)[a]	x				x							
Lee *et al.*, (2016)[b]	x				x							
Lee *et al.*, (2016)[c]	x				x							
Lee *et al.*, (2016)[d]	x				x							
Lee *et al.*, (2016)[e]	x				x							
Lee *et al.*, (2016)[f]	x				x							
Dewhurst *et al.*, (2014)	x				x							
Giné-Garriga *et al.*, (2010)[a]							x					
Giné-Garriga *et al.*, (2010)[b]							x					
TSLC												
Collado-Mateo *et al.*, (2019)	x				x							
Chan *et al.*, (2016)							x					
Le Berre *et al.*, (2016)	x				x						x	
TVC												
Fernandez-Huerta et al., (2019)[a]	x				x							
Fernandez-Huerta et al., (2019)[b]	x				x							

Tabela 3 - Cont.

Testes	Confiabilidade			Erro de Medição			Validade de Critério			Teste de Hipóteses para Validade de Construto		
	+	?	-	+	?	-	+	?	-	+	?	-
Fernandez-Huerta et al., (2019)[c]	x				x							
Forte *et al.*, (2021)[a]	x											
Forte *et al.*, (2021)[b]	x											
Forte *et al.*, (2021)[c]	x				x							
Forte *et al.*, (2021)[d]	x											
Forte *et al.*, (2021)[e]	x											
Forte *et al.*, (2021)[f]			x		x							
Dewhurst *et al.*, (2014)	x				x							
Adell *et al.*, (2013)	x				x							
Giné-Garriga *et al.*, (2010)[a]							x					
Giné-Garriga *et al.*, (2010)[b]									x			
Giné-Garriga *et al.*, (2010)[c]							x					
Giné-Garriga *et al.*, (2010)[d]									x			
Sayers *et al.*, (2006)[a]							x					
Sayers *et al.*, (2006)[b]							x					
Sayers *et al.*, (2006)[c]									x			
Özden *et al.*, (2022)[a]	x				x							
Özden *et al.*, (2022)[b]									x			
Özden *et al.*, (2022)[c]									x			
Özden *et al.*, (2022)[d]	x				x							
Özden *et al.*, (2022)[e]									x			
Özden *et al.*, (2022)[f]									x			
Criss *et al.*, (2023)[a]	x											

Tabela 3 - Cont.

Testes	Confiabilidade			Erro de Medição			Validade de Critério			Teste de Hipóteses para Validade de Construto		
	+	?	-	+	?	-	+	?	-	+	?	-
Criss *et al.*, (2023)[b]	x				x							
Criss *et al.*, (2023)[c]									x			
Criss *et al.*, (2023)[d]									x			
Criss *et al.*, (2023)[e]									x			
Criss *et al.*, (2023)[f]							x					
Criss *et al.*, (2023)[g]									x			
Criss *et al.*, (2023)[h]									x			
Saito *et al.*, (2022)[a]	x											
Saito *et al.*, (2022)[b]							x					
Saito *et al.*, (2022)[c]									x			
TPMI												
Balachandran *et al.*, (2021)[a]	x				x					x		
Sherwood *et al.*, (2020)[a]	x											
Sherwood *et al.*, (2020)[b]	x											
Sherwood *et al.*, (2020)[c]							x					
Sherwood *et al.*, (2020)[d]							x					
Cruvinel-Cabral *et al.*, (2018)	x				x		x					
Giné-Garriga *et al.*, (2010)[a]									x			
Giné-Garriga *et al.*, (2010)[b]									x			
Giné-Garriga *et al.*, (2010)[c]									x			
Giné-Garriga *et al.*, (2010)[d]									x			
Giné-Garriga *et al.*, (2010)[e]									x			
Giné-Garriga *et al.*, (2010)[f]									x			

Nota: (+) suficiente, (?) indeterminado e (-) insuficiente. Fonte: Elaborado pelo autor.

4.5 Propriedades de medição sumarizadas por teste e qualidade geral das evidências

4.5.1 *TUG* - Sumarização e qualidade geral das evidências (*GRADE*)

Para a confiabilidade, o resultado sumarizado das 13 medidas, tiveram 12 suficientes (+) pelas consistências principalmente quando foram avaliadas pelo teste-reteste entre ensaios que foi a medida mais realizada seguida das medidas entre avaliadores, intra dia e correlação do *TUG* cronometrado manual e automaticamente, no entanto, em pessoas idosas institucionalizadas só houve medidas de teste-reteste. Entretanto, quando Collado-Mateo *et al.*, (2019), avaliaram a correlação entre *TUG* medido por cronômetro manual e portátil, apesar da alta correlação (0.979), o resultado foi baseado no *Rho* de Spearman ou *r* de Pearson quando apropriados, como já mencionado, essas medidas são consideradas inadequadas para este fim, e, portanto, foi avaliada como indeterminada (?). A qualidade geral da evidência foi rebaixada em um nível, pois na análise do risco de viés, dos seis estudos correspondes as propriedades de medição confiabilidade, os estudos de Collado-Mateo *et al.*, (2019) e Dewhurst; Bampouras (2014), foram avaliados com qualidade metodológica duvidosa pelo *COSMIN Risk of Bias Checklist*, resultando em uma avaliação moderada pelo *GRADE*.

A avaliação da validade de critério teve como resultado sumarizado suficiente (+), pois as medidas de resultados foram consistentes quando foi avaliada a validade concorrente entre o *TUG* medido por um aplicativo de smartphone e medido por um sensor de força considerado como medida de referência (CHAN *et al.*, 2016), e entre o *TUG* modificado com dupla tarefa (cognitiva e motora) e TVC normal e rápida somente em pessoas idosas da comunidade (GINÉ-GARRIGA *et al.*, 2010). A qualidade geral da evidência foi rebaixada em um nível pela imprecisão, pois envolveu apenas dois estudos com amostras inferiores a 100, resultando em uma avaliação moderada pelo *GRADE*.

Já na avaliação do teste de hipóteses para validade de construto, o estudo de Le Berre *et al.*, (2016), descreveu ter realizado a validade convergente entre o *TUG* e o TSLC de 30 segundos modificado com o auxílio dos braços em pessoas idosas institucionalizadas, no entanto, os autores não descreveram nenhuma hipótese prévia, e, portanto, sua avaliação foi indeterminada (?) seguindo os critérios *COSMIN* para as boas propriedades de medição. Sendo assim, a sumarização se mantém indeterminado (?) e o *GRADE* é não aplicável (n/a) bem como o erro de medição. Todos os resultados das propriedades de medição sumarizadas para o *TUG* e a qualidade geral das evidências podem ser vistas na tabela 4.

4.5.2 TSLC - Sumarização e qualidade geral das evidências (*GRADE*)

A avaliação do domínio confiabilidade, resultou na sumarização suficiente (+), as medidas de resultados foram consistentes para a confiabilidade teste-reteste envolvendo o TSLC de 30 segundos medido automaticamente em pessoas idosas da comunidade (COLLADO-MATEO *et al.*, 2019) e o TSLC 30 s modificado com o uso dos braços em pessoas idosas institucionalizadas (LE BERRE *et al.*, 2016). A qualidade geral da evidência foi rebaixada em um nível na análise do risco de viés, pois dos dois estudos avaliados para a confiabilidade, o estudo de Collado-Mateo *et al.*, (2019), foi avaliado com qualidade metodológica duvidosa pelo *COSMIN Risk of Bias Checklist*, e por esse estudo ter a maior amostra entre os dois, decidiu-se por rebaixá-lo em mais um nível pela imprecisão, resultando em uma avaliação baixa pelo *GRADE*.

Para a validade de critério, o resultado sumarizado foi suficiente (+), as medidas de resultados foram consistentes quando foi avaliada a validade concorrente entre o TSLC 5 vezes medido por um aplicativo para smartphone e o TSLC 5 vezes medido por um sensor de força considerado como medida de referência em pessoas idosas da comunidade (CHAN *et al.*, 2016). A qualidade geral da evidência foi rebaixada em dois níveis pela imprecisão, o estudo tem baixo poder amostral (n= < 50), e por isso, o resultado da avaliação do *GRADE* é baixa.

Na avaliação do teste de hipóteses para validade de construto, como já mencionado, o estudo de Le Berre *et al.*, (2016), descreveu ter realizado a validade convergente entre o TSLC de 30 segundos modificado com o auxílio dos braços e o *TUG*, no entanto, os autores não descreveram nenhuma hipótese prévia, e, portanto, sua avaliação foi indeterminado (?) seguindo os critérios *COSMIN* para as boas propriedades de medição. Sendo assim, a sumarização se mantém indeterminado (?) e o *GRADE* é não aplicável (n/a) bem como o erro de medição. Todos os resultados das propriedades de medição sumarizadas para o TSLC e a qualidade geral das evidências podem ser vistas na tabela 4.

4.5.3 TVC - Sumarização e qualidade geral das evidências (*GRADE*)

Para a confiabilidade, com exceção do estudo de Forte *et al*., (2021), na confiabilidade entre sessões (*ICC*= 0.628; 95% *CI*= 0.431-0.768) a avaliação foi insuficiente (-) seguindo os critérios *COSMIN* para as boas propriedades de medição, no entanto, todas as medidas de resultados restantes foram consistentes principalmente quando foram avaliadas pelo teste-reteste entre ensaios (ADELL; WEHMHÖRNER; RYDWIK, 2013; CRISS *et al.*, 2023; FERNÁNDEZ-HUERTA; CÓRDOVA-LEÓN, 2019; ÖZDEN *et al.*, 2022; SAITO *et al.*, 2022) seguida das medidas intra e entre sessões (DEWHURST; BAMPOURAS, 2014; FORTE;

DE VITO; BOREHAM, 2021) envolvendo velocidades lentas, habituais e máximas de caminhada para trás e para frente entre distâncias que variam de 4 a 400 m medido digitalmente por cronômetro e aplicativo de smartphone, além da medida de correlação entre o TVC de 4 m e o TVC de 10 m somente em pessoas idosas da comunidade resultando em uma avaliação sumarizado de suficiente (+).

Na avaliação da qualidade geral da evidência, houve rebaixamento em um nível, pois na análise do risco de viés, dos sete estudos correspondes as propriedades de medição confiabilidade, quatro (CRISS *et al.*, 2023; DEWHURST; BAMPOURAS, 2014; FERNÁNDEZ-HUERTA; CÓRDOVA-LEÓN, 2019; SAITO *et al.*, 2022) foram avaliados com qualidade metodológica duvidosa pelo *COSMIN Risk of Bias Checklist*, e como estes estudos estão entre os com maiores amostras, decidiu-se em rebaixar mais um nível pela imprecisão, resultando em uma avaliação baixa pelo *GRADE*.

Para a validade de critério, o resultado sumarizado foi inconsistente (±) devido algumas inconsistências das medidas de resultados oriundas principalmente da validade concorrente entre o TVC com o TSLC 5 vezes e entre o TVC em diferentes velocidades (CRISS *et al.*, 2023; ÖZDEN *et al.*, 2022; SAITO *et al.*, 2022; SAYERS *et al.*, 2006). Ou seja, a depender da velocidade que o TVC é realizado (lenta, habitual ou auto selecionada e máxima) o construto medido por ambos pode convergir ou divergir-se. Pois o TVC tem a velocidade como principal construto medido, e o TSLC 5 vezes a potência de membros inferiores.

Isso pode ser corroborado pelo estudo de Criss *et al.*, (2023), onde diferentes velocidades do TVC correlacionaram-se diferentemente com o TSLC 5 vezes (TVC lenta *vs. 5xSTS* = .082; TVC auto selecionada *vs*. *5xSTS*= .-564; TVC máxima *vs*. *5xSTS*= .-691) nota-se que ao passo que a velocidade do TVC aumenta, a correlação entre os testes aumenta também, sugerindo uma diminuição da divergência entre construtos, uma vez que a potência é o produto da velocidade multiplicada pela força.

Adicionalmente, as velocidades do TVC (lenta, habitual ou auto selecionada e máxima) parecem sofrer divergências quando correlacionadas (TVC lenta *vs.* TVC auto selecionada= .186; TVC lenta *vs.* TVC máxima= 0.50; TVC máxima *vs.* TVC auto selecionada= .783) nota-se que a correlação aumenta ao passo que a velocidade aumenta, esse é mais um ponto que corrobora com o exposto anterior (CRISS *et al.*, 2023). Quando o TVC é realizado com dupla tarefa tanto cognitiva como motora, parece alterar a velocidade principalmente quando realizada de forma rápida (TVC de 8 m com dupla tarefa cognitiva (*Kick-8 m*) *vs*. TVC habitual: *r*= 0.776; *Kick-8 m vs.* TVC rápida: *r*= 0.686; TVC de 8 m com dupla tarefa motora (*TT-kick 8 m*) *vs.* TVC habitual: *r*= 0.736; *TT-kick 8 m vs*. TVC rápida: *r*= 0.660) mais um ponto

corroborando que o aumento da velocidade parece impactar o construto no qual o TVC se propõe a medir (GINÉ-GARRIGA *et al.*, 2010).

Todas as divergências nas correlações dos tipos de TVC com o TSLC 5 vezes e entre o TVC em diferentes velocidades, foram os principais geradores de inconsistências nas medidas de resultados para validade de critério, vide que, das 19 medidas de resultados consideradas para esta propriedade de medição, 7 são da validade concorrente entre o TVC e TSLC 5 vezes e 7 são da validade concorrente entre o TVC com diferentes velocidades. Além do TSLC 5 vezes, o *TUG* não teve uma boa correlação com o TVC de 3 m para trás e TVC de 12.24 m para frente, ambos tiveram uma correlação avaliada para validade de critério como insuficiente (-) de acordo com os critérios *COSMIN* para as boas propriedades de medição (ÖZDEN *et al.*, 2022).

No entanto, o TVC obteve boas correlações na validade concorrente com outros TVC realizados com a mesma ou parecida velocidade de execução (TVC habitual *vs.* TVC rápida), entre distâncias diferentes (400 m *vs.* 4 m), entre bateria de testes de desempenho físico funcional (*400 m W vs. SPPB*) e entre participantes do estudo e instrutores fitness do estudo (SAITO *et al.*, 2022; GINÉ-GARRIGA *et al.*, 2010; SAYERS *et al.*, 2006). Sendo assim, a qualidade geral da evidência foi rebaixada em dois níveis pelas inconsistências dos resultados principalmente na divergência entre construtos medidos, resultando em uma avaliação baixa pelo *GRADE*. Todos os resultados das propriedades de medição sumarizadas para o TVC e a qualidade geral das evidências podem ser vistas na tabela 4.

4.5.4 TPMI - Sumarização e qualidade geral das evidências (*GRADE*)

Para a confiabilidade, o resultado sumarizado foi suficiente (+), as medidas de resultados foram consistentes para a confiabilidade testes-reteste para o *STSp* 3 vezes medido por um *LPT* em pessoas idosas da comunidade, para confiabilidade entre avaliadores do TSLC de 5 vezes medido por análise de videografia 2D e do *CMJ* medido pelo aplicativo para smartphone *my jump* em pessoas idosas da comunidade e institucionalizadas respectivamente.

A qualidade geral da evidência para confiabilidade foi rebaixada em dois níveis na análise do risco de viés, dos três estudos correspondes as propriedades de medição confiabilidade, o estudo de Cruvinel-Cabral *et al.*, (2018), foi avaliado com qualidade metodológica inadequada pelo *COSMIN Risk of Bias Checklist*, particularmente em relação a administração das medidas nos participantes, sendo considerado com risco de viés muito sério, resultando em uma avaliação baixa pelo *GRADE*.

Para a validade de critério, o resultado sumarizado foi suficiente (+), embora as medidas de resultados tenham tido inconsistências devido a validade concorrente consideradas do estudo de Giné-Garriga *et al.*, (2010), oriundas do *TGUG* modicado com dupla tarefa, contendo duas fases que avaliam como construto a força dos membros inferiores correspondendo ao TPMI. Uma é o *SU*, e a outra *BT*, ambas avaliaram a validade concorrente com o TSLC de 5 vezes e a contração isométrica voluntária máxima (*MVC*) dos extensores dos joelhos direito e esquerdo em pessoas idosas da comunidade. Todas as correlações foram avaliadas com insuficiente (-) seguindo os critérios *COSMIN* para as boas propriedades de medição.

No entanto, as outras medidas de resultados dos outros dois estudos avaliados para esta propriedade de medição, foram consistentes para a validade concorrente entre o TSLC 5 vezes medido por análise de videografia 2D e o TSLC 5 vezes medido por *LPT* tanto para potência como para velocidade em pessoas idosas da comunidade (SHERWOOD *et al.*, 2020). E a validade concorrente entre o *CMJ* medido pelo aplicativo *My Jump* e o *CMJ* medido pelo tapete de contato em pessoas idosas da comunidade e institucionalizadas (CRUVINEL-CABRAL *et al.*, 2018).

Neste caso de inconsistência entre os estudos que avaliam a mesma propriedade de medição, os estudos com maior qualidade têm um maior peso na sumarização, assim como os estudos mais recentes em comparação com os mais antigos (MOKKINK, *et al.*, 2018; PRINSEN, *et al.*, 2018; TERWEE, *et al.*, 2018). Pela inconsistência dos resultados do estudo de Giné-Garriga *et al.*, (2010), a qualidade geral da evidência foi rebaixada em um nível resultando em uma avaliação moderada pelo *GRADE*.

Para avaliação do teste de hipóteses para a validade de construto, somente um estudo correspondeu para esta propriedade de medição. O estudo de Balachandran *et al.*, (2021), pré-definiu suas hipóteses para a validade convergente entre o *STSp* de 3 vezes medido por um *LPT* e um *LP* pneumático, *SPPB* que inclui um teste de equilíbrio, velocidade habitual de caminhada e testes de levantar da cadeira; *TUG* em ritmo normal e rápido e *PROMs* por meio de dois questionários (questionário de mobilidade e função física) e para avaliar a validade discriminante, esperava-se que o pico de potência do *STSp* seria menor nas mulheres do que nos homens.

Como hipotetizado, a potência de pico do *STSp* apresentou alta correlação de 0.90 com a potência do *LP*. Para resultados de desempenho físico, a potência de pico do *STSp* apresentou correlações (*r* de Pearson) semelhantes ou superiores (0,05) com os testes de função física em relação à potência de pico do *LP*: *SPPB* (0,41 *vs*. 0,29), TSLC 5 vezes (-0,44 *vs*. -0,35), *TUG* em ritmo habitual (-0,37 *vs*. -0,29) e ritmo rápido (-0,41 *vs*. -0,34) e equilíbrio (0,33 *vs*. 0,22)

mas não para o questionário de mobilidade (0,34 vs. 0,38) e função (0,41 *vs*. 0,48). Para a validade discriminante, como esperado, os homens apresentaram maior potência de pico no *STSp* em comparação com as mulheres (Δ= 492 W, p <0,001, d de Cohen= 2.0). Como pode ser observado, além das medidas descritas acima, há mais uma medida de confiabilidade teste-reteste do *STSp* mencionada no tópico anterior, totalizando dez medidas de resultados, no entanto, a avaliação da validade de construto se baseou na construção das hipóteses desenvolvidas, e, portanto, todas as medidas geraram uma avaliação geral.

Sendo assim, com mais de 75% das hipóteses confirmadas e as medidas de resultados avaliadas como suficientes (+) na avaliação geral, o resultado sumarizado do teste de hipóteses para validade de construto também foi considerado suficiente (+). No entanto, a qualidade geral da evidência foi rebaixada em um nível devido à indiretividade, com algumas divergências entre construtos resultando em correlações fracas, e mais um nível devido à imprecisão, dado o tamanho da amostra inferior a 100, resultando em uma avaliação baixa pelo *GRADE*. Todos os resultados das propriedades de medição sumarizadas para o TPMI e a qualidade geral das evidências estão disponíveis na tabela 4.

Tabela 4 - Avaliação sumarizada e *GRADE* das evidências para as propriedades de medição

	TUG		Teste Sentar e levantar da cadeira		Teste de Velocidade de caminhada		Teste de Potência de membros inferiores	
	Avaliação Geral	GRADE	Avaliação Geral	GRADE	Avaliação Geral	GRADE	Avaliação Geral	GRADE
CON	+	Moderada	+	Baixa	+	Baixa	+	Baixa
EM	?	n/a	?	n/a	?	n/a	?	n/a
VC	+	Moderada	+	Baixa	±	Baixa	+	Moderada
THV	?	n/a	?	n/a	n/a	n/a	+	Baixa

Nota: CON= confiabilidade, EM= erro de medição, VC= validade de critério, THV= teste de hipóteses para validade de construto, suficiente (+), inconsistente (±), indeterminado (?) e não aplicável (n/a). Fonte: Elaborado pelo autor.

5 DISCUSÃO

Considerando o que se tem conhecimento, é possível que esta seja a primeira revisão sistemática dedicada à avaliação das propriedades de medição de testes de medidas objetivas centradas no desempenho, destinadas a mensurar a função física funcional, com enfoque nos membros inferiores (*TUG*, TSLC, TVC e TPMI) em pessoas idosas, tanto residentes da comunidade quanto institucionalizadas. A relevância desses testes, reconhecidos por sua praticidade e acessibilidade, é amplamente documentada na literatura, especialmente quando se trata de avaliar o declínio dos componentes da função física e funcional em pessoas idosas. Eles desempenham um papel crucial no rastreamento de síndromes geriátricas, como fragilidade e sarcopenia, como evidenciado por estudos recentes (BAEK *et al.*, 2023; BHASIN *et al.*, 2020; CHEN *et al.*, 2020; CRUZ-JENTOFT *et al.*, 2019; DENT *et al.*, 2018, 2019; ZANKER *et al.*, 2022). Além disso, esses instrumentos demonstram uma promissora capacidade de predição de desfechos clínicos significativos, incluindo quedas, fraturas, hospitalizações, institucionalização e mortalidade, conforme apontam diversos estudos (ABELLAN VAN KAN *et al.*, 2009; BARRY *et al.*, 2014; BEAUCHET *et al.*, 2011; CAVANAUGH *et al.*, 2018; HARVEY *et al.*, 2018; LANDI *et al.*, 2010; PEEL; KUYS; KLEIN, 2013; PUA; MATCHAR, 2019; SCHOENE *et al.*, 2013; ZANKER *et al.*, 2023).

Nenhum dos quatro tipos de testes avaliados atendeu a todas as propriedades de medição, conforme os critérios desta revisão e, por conseguinte, de acordo com os padrões de qualidade estabelecidos pela *COSMIN methodology for systematic reviews*. Os testes, de modo geral, demonstraram boa confiabilidade teste-reteste entre dois ensaios, confiabilidade intra e entre avaliadores, ou intra e entre dias e/ou sessão, utilizando desde o cronômetro digital comum até medidas automáticas com recursos mais tecnológicos para mensuração (ver detalhes nas tabelas 1 e 3). Entretanto, a propriedade de medição "erro de medição", crucial na avaliação da confiabilidade ao identificar o erro sistemático e aleatório que não decorre de mudanças reais no construto a ser mensurado, não satisfez nenhum dos critérios devido à ausência do *MIC*. Em virtude disso, não foi possível incluir essa propriedade na sumarização e na avaliação geral da qualidade das evidências, evidenciando a necessidade premente de esforços adicionais nessa área de medição.

No que tange à validade de critério, observa-se considerável confusão, especialmente entre às terminologias utilizadas e as taxonomias *COSMIN*. Os estudos conduzidos por (CHAN *et al*., 2016; CRISS *et al*., 2023; GINÉ-GARRIGA *et al*., 2010; ÖZDEN *et al*., 2022; SAITO *et al*., 2022; SAYERS *et al*., 2006), alegaram ter realizado a validade concorrente, no entanto, parece que, na prática, foram realizadas correlações aleatórias com outros testes, sem critérios

estabelecidos e com construtos divergentes. Essa abordagem pode ter contribuído para as inconsistências, sobretudo no TVC, quando correlacionado com o TSLC 5 vezes. Em contrapartida, nos demais testes (*TUG*, TSLC e TPMI), nos quais a validade concorrente foi conduzida com critérios e construtos adequadamente estabelecidos, observou-se uma maior consistência nos resultados (CHAN *et al*., 2016; COLLADO-MATEO *et al*., 2019; CRUVINEL-CABRAL *et al*., 2018; GINÉ-GARRIGA *et al*., 2010; LEE *et al*., 2016; NEPAL; BASAULA; SHARMA, 2020).

Na avaliação do teste de hipóteses, observou-se uma situação semelhante, com apenas dois estudos abordando essa propriedade de medição. Esses estudos relataram ter realizado a validade convergente, um método para confirmar a validade de construto por meio do teste de hipóteses. No estudo de Le Berre *et al*., (2016), foi correlacionado o *TUG* com o TSLC de 30 segundos, sem o estabelecimento prévio de hipóteses. Já no estudo de Balachandran *et al*., (2021), foi realizada a validade convergente entre o *STSp*, *TUG* normal e rápido, e TSLC 5 vezes. Com exceção do TSLC 5 vezes, que pode ser considerado o mesmo teste realizado de maneira diferente, apresentando construtos convergentes, observou-se uma correlação curiosamente não tão forte com o *STSp* (*r*= -0,44; (95% *CI* [-0,62, -0,12]). Já o *TUG* realizado em velocidade normal (*r*= 0.37, 95% *CI* [-0.57, -0.05]) e velocidade rápida (*r*= -0.41 (95% *CI* [-0.56, -0.14]) conforme já mencionado, a divergência entre construtos parece impactar na correlação entre testes, ressaltando a importância de critérios e construtos adequados para evitar confusões entre validade concorrente e convergente.

Diante desse cenário, torna-se uma preocupação legítima que muitos dos instrumentos amplamente adotados ainda não tenham sido devidamente validados, levantando dúvidas substanciais sobre sua confiabilidade e validade. No entanto, este resultado deve ser interpretado com cautela, uma vez que a qualidade da evidência para cada teste e as respectivas propriedades de medição foram moderadas e baixas, suscitando mais esforços para pesquisas futuras nesta área. Todavia, é pertinente compreender que ainda é possível empregar os testes mencionados, embora com algumas ressalvas, como será elucidado de maneira prática.

Um destaque inicial pode ser o *TUG* e o TSLC 5 vezes como testes alternativos para avaliar a capacidade funcional em pessoas idosas, comparando-os com um padrão de referência (Chan *et al*., 2016). Nesse cenário, um aplicativo de smartphone, baseado em dados provenientes de uma unidade inercial tridimensional (*IMU*) no sistema Android, foi empregado para medir a mobilidade funcional, no *TUG* foi utilizada ação de levantar-se de uma cadeira com os braços a 46 cm de altura, caminhar 3 m em uma velocidade de caminhada confortável

e individualizada e sentar-se, enquanto o TSLC 5 vezes consistiu em repetições de sentar e levantar o mais rápido possível em uma cadeira sem braços com 43 cm de altura.

O estudo supracitado adotou uma metodologia meticulosa, utilizando um sensor de força como referência durante as medições. Os resultados revelaram uma consistência notável entre as medições obtidas pelo smartphone e pelo sensor de força, indicando um *ICC* de 0,946 (95% *CI*, 0.889 - 0.973) para o *TUG* e 0,988 (95% *CI*, 0,976 - 0,994) para o TSLC 5 vezes. Esses dados destacam a confiabilidade desses testes quando realizados por meio de um smartphone, uma ferramenta prática e acessível.

Entretanto, a análise de Bland-Altman identificou um viés positivo, indicando uma discrepância entre as medições. Esse viés foi quantificado em 0,27 segundos (95% *LoA*, 1,66 s a 2,63 s) para o *TUG* e 0,48 segundos (95% *LoA*, 1,66 s a 2,63 s) para o TSLC 5 vezes. Os pesquisadores postulam que essa superestimação na medição do *TUG* pode ser atribuída ao intervalo entre o sinal sonoro e o início efetivo do movimento. Esses resultados destacam não apenas a eficácia do smartphone como uma ferramenta de medição para testes baseados em desempenho, mas também ressaltam a importância de abordar possíveis fontes de viés para garantir a precisão desses instrumentos na avaliação da capacidade funcional em pessoas idosas da comunidade.

Similarmente, o estudo de Cruvinel-Cabral *et al.*, (2018), identificou um meio alternativo prático e acessível para avaliar a capacidade de salto, relacionada à força muscular, potência, velocidade e amplitude dos movimentos dos membros inferiores. A pesquisa investigou a validade e confiabilidade do aplicativo móvel "*My Jump*" na avaliação da altura do salto vertical em pessoas idosas. Os participantes realizaram três *CMJ*, avaliados tanto por um tapete de contato quanto pelo aplicativo *My Jump*. Os resultados mostraram uma correlação quase perfeita entre os métodos ($r = 0.999$; $p = 0.000$), com uma boa concordância satisfatória (*LoA*= -0,5177 a 0,3255 cm). A altura média do *CMJ* para o salto mais alto foi de 10,78 ± 5,23 cm com o tapete de contato e 10,87 ± 5,32 cm com o *My Jump App*, apresentando um viés sistemático mínimo de 0,096 cm ($p = 0,007$).

Na análise de confiabilidade, o *ICC* do *My Jump App* foi de 0,948, a média do erro típico de medição (*TEM*) foi de 1,150 cm, e o coeficiente de variação (*CV*) foi de 10,10%. Esses resultados indicam que o *My Jump App* é uma ferramenta válida e confiável em comparação com o tapete de contato para avaliar o desempenho do salto vertical tanto em pessoas idosas da comunidade quanto em institucionalizadas. Esse achado é significativo, destacando a viabilidade do uso de tecnologias simples e acessíveis para avaliações clínicas em, oferecendo uma alternativa prática ao equipamento convencional.

Dois testes adicionais que se destacaram como instrumentos complementares na avaliação do desempenho físico funcional em pessoas idosas estão relacionados ao TVC. O primeiro deles originou-se do estudo conduzido por Sayers *et al*., (2006), trata-se de um TVC individualizada (*400 m W*), o qual avaliou a validade concorrente em relação ao amplamente utilizado *SPPB*. Os resultados revelaram correlações moderadas entre a pontuação geral do *SPPB* e a velocidade do *400 m W* (r= 0,74; p <0,001). Notavelmente, o tempo de desempenho no *400 m W* diferenciou grupos de pessoas idosas de alto e baixo desempenho em vários parâmetros de saúde, destacando uma discordância entre o desempenho dos *400 m W* e do *SPPB* em homens e mulheres idosos com alto funcionamento.

Ou seja, pessoas idosas com alto funcionamento e tempos lentos no *400 m W* apresentaram menor potência muscular e velocidade de contração mais lenta, mais condições médicas, usaram mais medicamentos e relataram mais quedas em comparação com aqueles com tempos mais rápidos no *400 m W*. Conclui-se que, mesmo em pessoas idosas aparentemente bem funcionais residentes da comunidade, o teste de caminhada de *400 m W* revelou limitações, sugerindo sua justificativa para avaliações mais abrangentes da capacidade funcional.

No entanto, é crucial enfatizar que as correlações entre o *400 m W* e os componentes do *SPPB* especificamente, as tarefas de equilíbrio (r= 0,31) e a elevação da cadeira (r= -0,53) demonstraram ser mais tênues em comparação com as correlações entre o *400 m W* e a velocidade de caminhada de 4 m (r= 0,84). Esse achado evidencia que a possível convergência entre os construtos do *400 m W* e o *SPPB* está, na realidade, entre o *400 m W* e o teste de velocidade de caminhada de 4 m. Isso reforça as preocupações e diretrizes previamente discutidas nesta revisão acerca da importância na escolha criteriosa dos testes para a avaliação da validade convergente na confirmação do teste de hipóteses para validade de construto. Neste contexto, surge a possibilidade de considerar o TVC de 4 m como uma alternativa mais prática em relação ao *SPPB* e ao próprio *400 m W*.

Outro aspecto relevante a ser ponderado é que, além da exigência de espaço para a realização do teste de *400 m W*, um contingente significativo de participantes (34 de um total de 101, sendo 23 mulheres e 11 homens) não conseguiu completar o teste. Esse dado sugere que o percurso extenso pode representar um desafio adicional, indicando a necessidade de considerações cuidadosas em relação à viabilidade e acessibilidade desse teste em particular.

O segundo teste é sugerido pelo estudo de Saito *et al.*, (2022), trata-se do TVC habitual de 10 m no qual empregou-se um smartphone equipado com o aplicativo índice *ME-BYO* a fim de mensurar a velocidade de caminhada. Este aplicativo, projetado para armazenar dados localmente, inclui o índice *ME-BYO* de uso geral, no aplicativo *My ME-BYO Record*, disponível

para download na *App Store*. Durante o TVC de 10 m, os participantes seguiram as instruções fornecidas na tela para percorrer essa distância em sua velocidade habitual, sendo o tempo medido através da função cronômetro do aplicativo. A precisão da medição foi assegurada, requerendo que os participantes confirmassem e repetissem a medição se a velocidade diferisse do habitual. O tempo foi registrado com precisão de centésimos de segundo, e a velocidade de caminhada foi calculada como a distância percorrida (10 m) dividida pelo tempo (m/s).

Os resultados da validade de critério da velocidade de caminhada medida pelos participantes utilizando o índice *ME-BYO* e comparados com instrutores de fitness revelaram uma correlação positivamente forte (r= 0,862, *CI* 95% = 0,753, 0,925, $p < 0,001$). A análise de Bland-Altman demonstrou elevada concordância entre profissionais e participantes, com um erro sistemático dentro da faixa aceitável. Quanto à confiabilidade teste-reteste, o *ICC* foi de 0,712 (*CI* 95% = 0,571, 0,823, $p < 0,001$), indicando uma confiabilidade moderada. Conclui-se que o índice *ME-BYO*, quando autenticado pelos próprios participantes pessoas idosas da comunidade, emerge como uma alternativa robusta e concordante em relação às medições realizadas por profissionais instrutores fitness.

Essa abordagem autônoma oferece uma série de vantagens, destacando-se a praticidade e conveniência para as pessoas idosas, permitindo que eles realizem suas próprias avaliações de velocidade de caminhada de maneira simples e eficiente. Além disso, a alta concordância entre as medições dos participantes e instrutores, aliada à confiabilidade moderada observada nos testes repetidos, reforça a viabilidade e a utilidade dessa abordagem autogerenciada. Essa autonomia pode promover uma maior adesão das pessoas idosas às avaliações de desempenho físico, contribuindo para uma monitorização contínua e acessível de sua saúde e funcionalidade.

Por fim, o recente estudo de Balachandran *et al*., (2021), a fim de estabelecer a validade de construto, geraram-se previsões com base na construção de hipóteses, e essas previsões foram testadas para dar apoio à validade do instrumento pelos métodos de validade convergente e validade discriminativa. Este estudo teve como objetivo avaliar *STSp* 3 vezes medido por um *LPT* como uma medida eficaz da potência da parte inferior do corpo em adultos com 65 anos ou mais da comunidade. Os resultados revelaram que a potência de pico do STSp apresentou uma forte correlação com o *leg press* pneumático (*LP*) (r= 0,90, 95% *CI* [0,82, - 0,94]). Como hipotetizado, o *STSp* mostrou correlações similares ou superiores com testes de função física em comparação com a potência de pico do *LP*, evidenciando sua validade convergente.

Ao analisar a relação do *STSp* com outros testes, verificou-se correlações equivalentes e até superiores em alguns casos. Em relação à potência de pico do *LP*, o *STSp* mostrou correlações mais fortes com o *SPPB* (0,41 *vs*. 0,29), TSLC 5 vezes (-0,44 *vs*. -0,35), *TUG* em

ritmo habitual (-0,37 *vs*. -0,29) e ritmo rápido (-0,41 *vs*. -0,34), e equilíbrio (0,33 *vs*. 0,22). Além disso, o *STSp* apresentou validade discriminante, destacando diferenças significativas na potência de pico entre homens e mulheres. Quanto à confiabilidade, o *STSp* foi submetido a um teste-reteste em 36 participantes, resultando em um *ICC*= 0,96 e um erro padrão de medida de 70,4 W. Esses resultados sugerem que o *STSp* é uma ferramenta válida, confiável e eficaz na avaliação da potência da parte inferior do corpo em pessoas idosas residentes na comunidade. Sua rapidez, acessibilidade e segurança reforçam sua aplicabilidade em avaliações sobre o envelhecimento.

Embora praticamente todos os testes e suas respectivas medidas de resultados avaliadas possam ter demonstrado, em uma quantidade restrita de estudos, uma confiabilidade aceitável de forma geral, a falta de avaliação do erro de medição exige prudência nessa análise, considerando que o erro de medição é uma propriedade essencial no domínio da confiabilidade.

Esta revisão tem vários pontos fortes. Não houve restrição de idioma na inclusão de estudos, mitigando assim o risco de viés de idioma (SHIWA *et al*., 2013). Além disso, aderiu rigorosamente às recomendações do manual *COSMIN methodology for systematic reviews* para estudos sobre propriedades de medição, incluindo a avaliação cuidadosa da qualidade das evidências, conforme as diretrizes da *Modified Grading of Recommendations Assessment, Development and Evaluation* (*GRADE*). Esta abordagem permite uma avaliação criteriosa da qualidade das evidências e, por conseguinte, da certeza dos resultados. Vale destacar também que foi empregada uma estratégia de busca altamente sensível para identificar estudos nas principais bases de dados, englobando tanto pessoas idosas da comunidade quanto institucionalizadas (MOKKINK, *et al*., 2018; PRINSEN, *et al*., 2018; TERWEE, *et al*., 2018).

No entanto, esta revisão apresenta algumas limitações. Uma delas é a escolha deliberada de um escopo mais restrito, com critérios de elegibilidade centrados na resposta a uma pergunta de pesquisa mais específica. Nesse sentido, foram incluídos apenas estudos que abordaram um dos quatro testes e que, em seus processos metodológicos de construção, adaptação e validação, não comprometessem a acessibilidade e praticidade para sua realização. Essa abordagem foi adotada para evitar a ampliação excessiva do escopo, o que poderia resultar em uma síntese de evidências abrangente, mas também em um aumento de inconsistências ao misturar vários métodos de mensuração e execução dos testes. Essa complexidade poderia dificultar a gestão da revisão, interpretação dos dados e clareza das conclusões. Vale ressaltar que, devido a essa abordagem, alguns estudos que utilizaram recursos tecnológicos considerados inacessíveis devido a altos custos ou materiais específicos foram excluídos da revisão.

Outra limitação potencial decorrente do escopo desta revisão é que muitos estudos não apresentaram de maneira clara seus processos de construção, adaptação e validação dos testes. Por exemplo, alguns estudos mencionavam realizar apenas uma validação, mas não especificavam o tipo de validação. Devido à confusão de terminologias e à semelhança de alguns processos entre as propriedades de medição já mencionadas, poderia haver equívocos ao tentar esclarecer a propriedade de medição avaliada pelo estudo, conforme estabelecido pelos critérios de inclusão (ii).

Além disso, o escopo da revisão concentrou-se em pessoas idosas considerados saudáveis, excluindo estudos nos quais os processos de construção, adaptação e validação foram conduzidos em populações de pessoas idosas com doenças específicas, o que limita ainda mais a generalização dos resultados. Dessa forma, é possível que os estudos excluídos pelos critérios de elegibilidade, orientados pelo escopo desta revisão, poderiam ter resultados diferentes se fossem incluídos. Contudo, é possível que alguns estudos tenham sido publicados em bases de dados locais e consequentemente não tenham sido encontrados e incluídos nesta revisão, o que pode ser considerado uma outa potencial limitação da revisão.

6 CONCLUSÃO

A importância dos testes de avaliação baseados no desempenho físico funcional é inegável, desempenhando um papel crucial na compreensão do estado de saúde e na orientação de decisões clínicas. Contudo, os resultados desta revisão destacam que muitos desses instrumentos carecem de evidências sólidas em relação às suas propriedades de medição, frequentemente apresentando limitações significativas.

Na prática, para a avaliação da capacidade funcional, especialmente em tarefas que exigem mobilidade funcional, o *TUG* e o TSLC 5 vezes com o uso de um aplicativo para smartphone, baseado em dados provenientes de uma *IMU* no sistema *Android,* o TVC de 10 m medido por um smartphone equipado com o índice *ME-BYO* no aplicativo *My ME-BYO Record* no sistema *IOS*, o TVC individualizada *400 m W*, manualmente cronometrado por um aparelho digital comum, e o TVC de 4 m pela maior praticidade e correlação com o *400 m W* é uma opção a considerar. Todas essas alternativas mostram vantagens consideráveis em termos de magnitude e consistência das propriedades de medição avaliadas, oferecendo flexibilidade na escolha de testes de acordo com as necessidades específicas da avaliação.

Para a avaliação da potência dos membros inferiores, a avaliação do salto vertical *CMJ* por meio do aplicativo *My Jump* para o sistema *IOS* e o *STSp* de 3 vezes medido por um *LPT* surgem como ferramentas para essa finalidade. Essas opções apresentam-se como alternativas viáveis e mais seguras, fornecendo informações confiáveis sobre a força e potência muscular dos membros inferiores em diferentes contextos de avaliação, permitindo uma abordagem mais abrangente e personalizada na escolha dos testes conforme as características e objetivos específicos de cada avaliação.

Para as pesquisas futuras, sugere-se que, os estudos sigam de preferência as diretrizes *COSMIN*, para uma padronização nas etapas de construção, adaptação e validação de seus instrumentos. O *COSMIN,* representa uma iniciativa de grande relevância internacional pelos inúmeros avanços proporcionados no campo da mensuração do estado de saúde, no qual o trabalho do grupo *COSMIN*, baseiam-se no consenso entre uma equipe multidisciplinar internacional de mais de 40 investigadores, metodologistas e clínicos com experiência nesta área. Esses avanços incluem a formulação consensual de definições para propriedades de medição, a elaboração de listas de verificação destinadas à avaliação do desenvolvimento de instrumentos e a criação da taxonomia de medição *COSMIN*.

Entretanto, destaca-se a necessidade de aprimoramento nas diretrizes *COSMIN*, especialmente no que diz respeito à avaliação de instrumentos físicos que empregam medidas objetivas, semelhantes aos testes desta revisão. Uma observação importante é que muitos

termos ainda mantêm uma forte associação com os *PROMs*. Ao analisar o *COSMIN Risk of Bias checklist*, especificamente nas caixas 6 confiabilidade e 7 erro de medição (apêndice E), os itens relacionados às variáveis e à escala de medição referem-se principalmente a variáveis categóricas, enquanto os testes de desempenho físico funcional frequentemente utilizam escores com variáveis numéricas (discretas ou contínuas). Isso pode criar lacuna na avaliação desses itens, sem orientações claras sobre como proceder nessa situação, tanto na etapa de avaliação individual quanto na etapa de avaliação sumarizada, que depende dos resultados do *COSMIN Risk of Bias checklist*. Portanto, sugere-se uma revisão e adaptação mais precisa dessas diretrizes para refletir as nuances específicas dos instrumentos físicos de medidas objetivas.

Outro ponto que suscita dúvidas relevantes refere-se à avaliação dos critérios para as boas propriedades de medição. Alguns estudos realizam a avaliação da confiabilidade entre sessão, intra avaliador, entre avaliadores e correlação entre testes, o que significa várias medidas do mesmo teste para a mesma propriedade de medição. No entanto, não está claro se cada medida deve ser avaliada individualmente e como proceder diante de divergências nos resultados decorrentes das avaliações individuais.

Com a melhoria dessas questões, há uma perspectiva significativa de aprimoramento nos processos metodológicos dos estudos, especialmente no que diz respeito às propriedades de medição dos testes avaliados. Isso não apenas facilitaria a condução de estudos mais robustos, mas também tornaria mais acessível a realização de futuras revisões que aderem às diretrizes *COSMIN*. A padronização metodológica aprimorada proporcionaria maior clareza e consistência nos procedimentos de avaliação, minimizando potenciais equívocos originados pela falta de uniformidade, interpretação de termos e lacunas nas orientações, como evidenciado nesta revisão.

A adoção dessas melhorias não só beneficiaria a qualidade intrínseca dos estudos individuais, mas também fortaleceria a base de evidências disponível para pesquisadores e profissionais da saúde. A padronização metodológica contribuiria para a confiabilidade e validade das medições, promovendo uma base mais sólida para análises comparativas e interpretações consistentes em estudos futuros. Assim, a implementação dessas melhorias não só otimizaria os estudos existentes, mas também abriria caminho para avanços mais robustos no campo da avaliação de propriedades de medição de testes de desempenho físico funcional.

7 REFERÊNCIAS

ABELLAN VAN KAN, G.; ROLLAND, Y.; ANDRIEU, S.; BAUER, J.; BEAUCHET, O.; BONNEFOY, M.; CESARI, M.; DONINI, L. M.; GILLETTE GUYONNET, S.; INZITARI, M.; NOURHASHEMI, F.; ONDER, G.; RITZ, P.; SALVA, A.; VISSER, M.; VELLAS, B. Gait speed at usual pace as a predictor of adverse outcomes in community-dwelling older people an International Academy on Nutrition and Aging (IANA) Task Force. **The Journal of Nutrition, Health & Aging**, v. 13, n. 10, p. 881–889, dez. 2009. https://doi.org/10.1007/s12603-009-0246-z.

ADELL, E.; WEHMHÖRNER, S.; RYDWIK, E. The Test-Retest Reliability of 10 Meters Maximal Walking Speed in Older People Living in a Residential Care Unit. **Journal of Geriatric Physical Therapy**, v. 36, n. 2, p. 74–77, abr. 2013. https://doi.org/10.1519/JPT.0b013e318264b8ed.

ANGULO, J.; EL ASSAR, M.; ÁLVAREZ-BUSTOS, A.; RODRÍGUEZ-MAÑAS, L. Physical activity and exercise: Strategies to manage frailty. **Redox Biology**, v. 35, p. 101513, 20 mar. 2020. https://doi.org/10.1016/j.redox.2020.101513.

BAEK, J. Y.; JUNG, H.-W.; KIM, K. M.; KIM, M.; PARK, C. Y.; LEE, K.-P.; LEE, S. Y.; JANG, I.-Y.; JEON, O. H.; LIM, J.-Y. Korean Working Group on Sarcopenia Guideline: Expert Consensus on Sarcopenia Screening and Diagnosis by the Korean Society of Sarcopenia, the Korean Society for Bone and Mineral Research, and the Korean Geriatrics Society. **Annals of Geriatric Medicine and Research**, v. 27, n. 1, p. 9–21, mar. 2023. https://doi.org/10.4235/agmr.23.0009.

BALACHANDRAN, A. T.; VIGOTSKY, A. D.; QUILES, N.; MOKKINK, L. B.; BELIO, M. A.; GLENN, J. M. Validity, reliability, and measurement error of a sit-to-stand power test in older adults: A pre-registered study. **Experimental Gerontology**, v. 145, p. 111202, mar. 2021. https://doi.org/10.1016/j.exger.2020.111202.

BARRY, E.; GALVIN, R.; KEOGH, C.; HORGAN, F.; FAHEY, T. Is the Timed Up and Go test a useful predictor of risk of falls in community dwelling older adults: a systematic review and meta-analysis. **BMC geriatrics**, v. 14, p. 14, 1 fev. 2014. https://doi.org/10.1186/1471-2318-14-14.

BEAUCHET, O.; FANTINO, B.; ALLALI, G.; MUIR, S. W.; MONTERO-ODASSO, M.; ANNWEILER, C. Timed Up and Go test and risk of falls in older adults: a systematic review. **The Journal of Nutrition, Health & Aging**, v. 15, n. 10, p. 933–938, dez. 2011. https://doi.org/10.1007/s12603-011-0062-0.

BHASIN, S.; TRAVISON, T. G.; MANINI, T. M.; PATEL, S.; PENCINA, K. M.; FIELDING, R. A.; MAGAZINER, J. M.; NEWMAN, A. B.; KIEL, D. P.; COOPER, C.; GURALNIK, J. M.; CAULEY, J. A.; ARAI, H.; CLARK, B. C.; LANDI, F.; SCHAAP, L. A.; PEREIRA, S. L.; ROOKS, D.; WOO, J.; … CAWTHON, P. M. Sarcopenia Definition: The Position Statements of the Sarcopenia Definition and Outcomes Consortium. **Journal of the American Geriatrics Society**, v. 68, n. 7, p. 1410–1418, jul. 2020. https://doi.org/10.1111/jgs.16372.

BOHANNON, R. W. Hand-grip dynamometry predicts future outcomes in aging adults. **Journal of Geriatric Physical Therapy (2001)**, v. 31, n. 1, p. 3–10, 2008. https://doi.org/10.1519/00139143-200831010-00002.

BRACH, J. S.; VANSWEARINGEN, J. M.; NEWMAN, A. B.; KRISKA, A. M. Identifying early decline of physical function in community-dwelling older women: performance-based and self-report measures. **Physical Therapy**, v. 82, n. 4, p. 320–328, abr. 2002.

BRANCH, L. G.; MEYERS, A. R. Assessing physical function in the elderly. **Clinics in Geriatric Medicine**, v. 3, n. 1, p. 29–51, fev. 1987.

BUCHNER, D. M.; LARSON, E. B.; WAGNER, E. H.; KOEPSELL, T. D.; DE LATEUR, B. J. Evidence for a non-linear relationship between leg strength and gait speed. **Age and Ageing**, v. 25, n. 5, p. 386–391, set. 1996. https://doi.org/10.1093/ageing/25.5.386.

CAVANAUGH, E. J.; RICHARDSON, J.; MCCALLUM, C. A.; WILHELM, M. The Predictive Validity of Physical Performance Measures in Determining Markers of Preclinical Disability in Community-Dwelling Middle-Aged and Older Adults: A Systematic Review. **Physical Therapy**, v. 98, n. 12, p. 1010–1021, 1 dez. 2018. https://doi.org/10.1093/ptj/pzy109.

CHAD, K. E.; REEDER, B. A.; HARRISON, E. L.; ASHWORTH, N. L.; SHEPPARD, S. M.; SCHULTZ, S. L.; BRUNER, B. G.; FISHER, K. L.; LAWSON, J. A. Profile of physical activity levels in community-dwelling older adults. **Medicine and Science in Sports and Exercise**, v. 37, n. 10, p. 1774–1784, out. 2005. https://doi.org/10.1249/01.mss.0000181303.51937.9c.

CHAN, M. H. M.; KEUNG, D. T. F.; LUI, S. Y. T.; CHEUNG, R. T. H. A validation study of a smartphone application for functional mobility assessment of the elderly. **Hong Kong Physiotherapy Journal**, v. 35, p. 1–4, dez. 2016. https://doi.org/10.1016/j.hkpj.2015.11.001.

CHEN, L.-K.; WOO, J.; ASSANTACHAI, P.; AUYEUNG, T.-W.; CHOU, M.-Y.; IIJIMA, K.; JANG, H. C.; KANG, L.; KIM, M.; KIM, S.; KOJIMA, T.; KUZUYA, M.; LEE, J. S. W.; LEE, S. Y.; LEE, W.-J.; LEE, Y.; LIANG, C.-K.; LIM, J.-Y.; LIM, W. S.; … ARAI, H. Asian Working Group for Sarcopenia: 2019 Consensus Update on Sarcopenia Diagnosis and Treatment. **Journal of the American Medical Directors Association**, v. 21, n. 3, p. 300-307.e2, mar. 2020. https://doi.org/10.1016/j.jamda.2019.12.012.

COLLADO-MATEO, D.; MADEIRA, P.; DOMINGUEZ-MUÑOZ, F. J.; VILLAFAINA, S.; TOMAS-CARUS, P.; PARRACA, J. A. The Automatic Assessment of Strength and Mobility in Older Adults: A Test-Retest Reliability Study. **Medicina**, v. 55, n. 6, p. 270, 11 jun. 2019. https://doi.org/10.3390/medicina55060270.

COOPER, R.; KUH, D.; HARDY, R.; MORTALITY REVIEW GROUP. Objectively measured physical capability levels and mortality: systematic review and meta-analysis. **The BMJ**, v. 341, p. c4467, 9 set. 2010. https://doi.org/10.1136/bmj.c4467.

CRISS, M. G.; CHUI, K. K.; GALLICHIO, J.; CENTRA, J.; CANBEK, J. Reliability, responsiveness, and validity of slow walking speed in community dwelling older adults. **Gait & Posture**, v. 99, p. 54–59, jan. 2023. https://doi.org/10.1016/j.gaitpost.2022.10.016.

CRUVINEL-CABRAL, R. M.; OLIVEIRA-SILVA, I.; MEDEIROS, A. R.; CLAUDINO, J. G.; JIMÉNEZ-REYES, P.; BOULLOSA, D. A. The validity and reliability of the "*My Jump App*" for measuring jump height of the elderly. **PeerJ**, v. 6, p. e5804, 15 out. 2018. https://doi.org/10.7717/peerj.5804.

CRUZ-JENTOFT, A. J.; BAHAT, G.; BAUER, J.; BOIRIE, Y.; BRUYÈRE, O.; CEDERHOLM, T.; COOPER, C.; LANDI, F.; ROLLAND, Y.; SAYER, A. A.; SCHNEIDER, S. M.; SIEBER, C. C.; TOPINKOVA, E.; VANDEWOUDE, M.; VISSER, M.; ZAMBONI, M.; WRITING GROUP FOR THE EUROPEAN WORKING GROUP ON SARCOPENIA IN OLDER PEOPLE 2 (EWGSOP2), AND THE EXTENDED GROUP FOR EWGSOP2. Sarcopenia: revised European consensus on definition and diagnosis. **Age and Ageing**, v. 48, n. 1, p. 16–31, 1 jan. 2019. https://doi.org/10.1093/ageing/afy169.

CSAPO, R.; GORMASZ, C.; BARON, R. Functional performance in community-dwelling and institutionalized elderly women. **Wiener Klinische Wochenschrift**, v. 121, n. 11–12, p. 383–390, 2009. https://doi.org/10.1007/s00508-009-1151-5.

CSUKA, M.; MCCARTY, D. J. Simple method for measurement of lower extremity muscle strength. **The American Journal of Medicine**, v. 78, n. 1, p. 77–81, jan. 1985. https://doi.org/10.1016/0002-9343(85)90465-6.

CURRELL, K.; JEUKENDRUP, A. E. Validity, reliability and sensitivity of measures of sporting performance. **Sports Medicine (Auckland, N.Z.)**, v. 38, n. 4, p. 297–316, 2008. https://doi.org/10.2165/00007256-200838040-00003.

DE MORTON, N. A.; BERLOWITZ, D. J.; KEATING, J. L. A systematic review of mobility instruments and their measurement properties for older acute medical patients. **Health and Quality of Life Outcomes**, v. 6, p. 44, 5 jun. 2008. https://doi.org/10.1186/1477-7525-6-44.

DENT, E.; MORLEY, J. E.; CRUZ-JENTOFT, A. J.; ARAI, H.; KRITCHEVSKY, S. B.; GURALNIK, J.; BAUER, J. M.; PAHOR, M.; CLARK, B. C.; CESARI, M.; RUIZ, J.; SIEBER, C. C.; AUBERTIN-LEHEUDRE, M.; WATERS, D. L.; VISVANATHAN, R.; LANDI, F.; VILLAREAL, D. T.; FIELDING, R.; WON, C. W.; … VELLAS, B. International Clinical Practice Guidelines for Sarcopenia (ICFSR): Screening, Diagnosis and Management. **The Journal of Nutrition, Health & Aging**, v. 22, n. 10, p. 1148–1161, 2018. https://doi.org/10.1007/s12603-018-1139-9.

DENT, E.; MORLEY, J. E.; CRUZ-JENTOFT, A. J.; WOODHOUSE, L.; RODRÍGUEZ-MAÑAS, L.; FRIED, L. P.; WOO, J.; APRAHAMIAN, I.; SANFORD, A.; LUNDY, J.; LANDI, F.; BEILBY, J.; MARTIN, F. C.; BAUER, J. M.; FERRUCCI, L.; MERCHANT, R. A.; DONG, B.; ARAI, H.; HOOGENDIJK, E. O.; … VELLAS, B. Physical Frailty: ICFSR International Clinical Practice Guidelines for Identification and Management. **The Journal of Nutrition, Health & Aging**, v. 23, n. 9, p. 771–787, 2019. https://doi.org/10.1007/s12603-019-1273-z.

DEWHURST, S.; BAMPOURAS, T. M. Intraday Reliability and Sensitivity of Four Functional Ability Tests in Older Women. **American Journal of Physical Medicine & Rehabilitation**, v. 93, n. 8, p. 703–707, ago. 2014. https://doi.org/10.1097/PHM.0000000000000078.

FEINSTEIN, A. R.; JOSEPHY, B. R.; WELLS, C. K. Scientific and clinical problems in indexes of functional disability. **Annals of Internal Medicine**, v. 105, n. 3, p. 413–420, set. 1986. https://doi.org/10.7326/0003-4819-105-3-413.

FERNÁNDEZ-HUERTA, L.; CÓRDOVA-LEÓN, K. Reliability of two gait speed tests of different timed phases and equal non-timed phases in community-dwelling older persons. **Medwave**, v. 19, n. 03, p. e7611–e7611, 15 abr. 2019. https://doi.org/10.5867/medwave.2019.03.7611.

FERRUCCI, L.; GURALNIK, J. M.; STUDENSKI, S.; FRIED, L. P.; CUTLER, G. B.; WALSTON, J. D.; INTERVENTIONS ON FRAILTY WORKING GROUP. Designing randomized, controlled trials aimed at preventing or delaying functional decline and disability in frail, older persons: a consensus report. **Journal of the American Geriatrics Society**, v. 52, n. 4, p. 625–634, abr. 2004. https://doi.org/10.1111/j.1532-5415.2004.52174.x.

FIEO, R. A.; AUSTIN, E. J.; STARR, J. M.; DEARY, I. J. Calibrating ADL-IADL scales to improve measurement accuracy and to extend the disability construct into the preclinical range: a systematic review. **BMC geriatrics**, v. 11, p. 42, 16 ago. 2011. https://doi.org/10.1186/1471-2318-11-42.

FIGUEREDO, D. J.; JACOB-FILHO, W. Comparison between subjective and objective evaluations of self-care performance in elderly inpatients. **Einstein**, v. 16, n. 1, 2018. DOI 10.1590/S1679-45082018AO3987. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6063749/. Acesso em: 28 ago. 2022.

FORTE, R.; DE VITO, G.; BOREHAM, C. A. G. Reliability of walking speed in basic and complex conditions in healthy, older community-dwelling individuals. **Aging Clinical and Experimental Research**, v. 33, n. 2, p. 311–317, fev. 2021. https://doi.org/10.1007/s40520-020-01543-x.

FREIBERGER, E.; DE VREEDE, P.; SCHOENE, D.; RYDWIK, E.; MUELLER, V.; FRÄNDIN, K.; HOPMAN-ROCK, M. Performance-based physical function in older community-dwelling persons: a systematic review of instruments. **Age and Ageing**, v. 41, n. 6, p. 712–721, nov. 2012. https://doi.org/10.1093/ageing/afs099.

GALHARDAS, L.; RAIMUNDO, A.; MARMELEIRA, J. Test-retest reliability of upper-limb proprioception and balance tests in older nursing home residents. **Archives of Gerontology and Geriatrics**, v. 89, p. 104079, jul. 2020. https://doi.org/10.1016/j.archger.2020.104079.

GINÉ-GARRIGA, M.; GUERRA, M.; MANINI, T. M.; MARÍ-DELL'OLMO, M.; PAGÈS, E.; UNNITHAN, V. B. Measuring balance, lower extremity strength and gait in the elderly: Construct validation of an instrument. **Archives of Gerontology and Geriatrics**, v. 51, n. 2, p. 199–204, set. 2010. https://doi.org/10.1016/j.archger.2009.10.008.

GURALNIK, J. M.; BRANCH, L. G.; CUMMINGS, S. R.; CURB, J. D. Physical performance measures in aging research. **Journal of Gerontology**, v. 44, n. 5, p. M141-146, set. 1989. https://doi.org/10.1093/geronj/44.5.m141.

GURALNIK, J. M.; SIMONSICK, E. M.; FERRUCCI, L.; GLYNN, R. J.; BERKMAN, L. F.; BLAZER, D. G.; SCHERR, P. A.; WALLACE, R. B. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. **Journal of Gerontology**, v. 49, n. 2, p. M85-94, mar. 1994. https://doi.org/10.1093/geronj/49.2.m85.

GUYATT, G. H.; THOMPSON, P. J.; BERMAN, L. B.; SULLIVAN, M. J.; TOWNSEND, M.; JONES, N. L.; PUGSLEY, S. O. How should we measure function in patients with chronic heart and lung disease? **Journal of Chronic Diseases**, v. 38, n. 6, p. 517–524, 1985. https://doi.org/10.1016/0021-9681(85)90035-9.

GUYATT, Gordon H.; OXMAN, A. D.; SCHÜNEMANN, H. J.; TUGWELL, P.; KNOTTNERUS, A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. **Journal of Clinical Epidemiology**, v. 64, n. 4, p. 380–382, abr. 2011. https://doi.org/10.1016/j.jclinepi.2010.09.011.

HARVEY, N. C.; ODÉN, A.; ORWOLL, E.; LAPIDUS, J.; KWOK, T.; KARLSSON, M. K.; ROSENGREN, B. E.; RIBOM, E.; COOPER, C.; CAWTHON, P. M.; KANIS, J. A.; OHLSSON, C.; MELLSTRÖM, D.; JOHANSSON, H.; MCCLOSKEY, E. Measures of Physical Performance and Muscle Strength as Predictors of Fracture Risk Independent of FRAX, Falls, and aBMD: A Meta-Analysis of the Osteoporotic Fractures in Men (MrOS) Study. **Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research**, v. 33, n. 12, p. 2150–2157, dez. 2018. https://doi.org/10.1002/jbmr.3556.

HAZUDA, H. P.; DHANDA, R.; OWEN, S. V.; LICHTENSTEIN, M. J. Development and validation of a performance-based measure of upper extremity functional limitation. **Aging Clinical and Experimental Research**, v. 17, n. 5, p. 394–401, out. 2005. https://doi.org/10.1007/BF03324629.

HOPMAN-ROCK, M.; VAN HIRTUM, H.; DE VREEDE, P.; FREIBERGER, E. Activities of daily living in older community-dwelling persons: a systematic review of psychometric properties of instruments. **Aging Clinical and Experimental Research**, v. 31, n. 7, p. 917–925, 2019. https://doi.org/10.1007/s40520-018-1034-6.

KATZ, S.; FORD, A. B.; MOSKOWITZ, R. W.; JACKSON, B. A.; JAFFE, M. W. STUDIES OF ILLNESS IN THE AGED. THE INDEX OF ADL: A STANDARDIZED MEASURE OF BIOLOGICAL AND PSYCHOSOCIAL FUNCTION. **JAMA**, v. 185, p. 914–919, 21 set. 1963. https://doi.org/10.1001/jama.1963.03060120024016.

KEMPEN, G. I.; VAN HEUVELEN, M. J.; VAN DEN BRINK, R. H.; KOOIJMAN, A. C.; KLEIN, M.; HOUX, P. J.; ORMEL, J. Factors affecting contrasting results between self-reported and performance-based levels of physical limitation. **Age and Ageing**, v. 25, n. 6, p. 458–464, nov. 1996. https://doi.org/10.1093/ageing/25.6.458.

KRUIANSKY, J.; GURLAND, B. The performance test of activities of daily living. **International Journal of Aging & Human Development**, v. 7, n. 4, p. 343–352, 1976. https://doi.org/10.2190/x45l-tww7-wxxy-ka6k.

LAMB, S. E.; KEENE, D. J. Measuring physical capacity and performance in older people. **Best Practice & Research Clinical Rheumatology**, v. 31, n. 2, p. 243–254, abr. 2017. https://doi.org/10.1016/j.berh.2017.11.008.

LANDI, F.; LIPEROTI, R.; RUSSO, A.; CAPOLUONGO, E.; BARILLARO, C.; PAHOR, M.; BERNABEI, R.; ONDER, G. Disability, more than multimorbidity, was predictive of mortality among older persons aged 80 years and older. **Journal of Clinical Epidemiology**, v. 63, n. 7, p. 752–759, jul. 2010. https://doi.org/10.1016/j.jclinepi.2009.09.007.

LAWTON, M. P.; BRODY, E. M. Assessment of older people: self-maintaining and instrumental activities of daily living. **The Gerontologist**, v. 9, n. 3, p. 179–186, 1969.

LE BERRE, M.; APAP, D.; BABCOCK, J.; BRAY, S.; GAREAU, E.; CHASSÉ, K.; LÉVESQUE, N.; ROBBINS, S. M. The Psychometric Properties of a Modified Sit-to-Stand Test With Use of the Upper Extremities in Institutionalized Older Adults. **Perceptual and Motor Skills**, v. 123, n. 1, p. 138–152, ago. 2016. https://doi.org/10.1177/0031512516653388.

LEE, S.-P.; DUFEK, J.; HICKMAN, R.; SCHUERMAN, S. Influence of Procedural Factors on the Reliability and Performance of the Timed Up-and-go Test in Older Adults. **International Journal of Gerontology**, v. 10, n. 1, p. 37–42, mar. 2016. https://doi.org/10.1016/j.ijge.2015.10.003.

LEÓN-SALAS, B.; AYALA, A.; BLAYA-NOVÁKOVÁ, V.; AVILA-VILLANUEVA, M.; RODRÍGUEZ-BLÁZQUEZ, C.; ROJO-PÉREZ, F.; FERNÁNDEZ-MAYORALAS, G.; MARTÍNEZ-MARTÍN, P.; FORJAZ, M. J.; SPANISH RESEARCH GROUP ON QUALITY OF LIFE AND AGING. Quality of life across three groups of older adults differing in cognitive status and place of residence. **Geriatrics & Gerontology International**, v. 15, n. 5, p. 627–635, maio 2015. https://doi.org/10.1111/ggi.12325.

LINDEMANN, U.; CLAUS, H.; STUBER, M.; AUGAT, P.; MUCHE, R.; NIKOLAUS, T.; BECKER, C. Measuring power during the sit-to-stand transfer. **European Journal of Applied Physiology**, v. 89, n. 5, p. 466–470, jun. 2003. https://doi.org/10.1007/s00421-003-0837-z.

MAGAZINER, J.; ZIMMERMAN, S. I.; GRUBER-BALDINI, A. L.; HEBEL, J. R.; FOX, K. M. Proxy reporting in five areas of functional status. Comparison with self-reports and observations of performance. **American Journal of Epidemiology**, v. 146, n. 5, p. 418–428, 1 set. 1997. https://doi.org/10.1093/oxfordjournals.aje.a009295.

MATHIAS, S.; NAYAK, U. S.; ISAACS, B. Balance in elderly patients: the “get-up and go” test. **Archives of Physical Medicine and Rehabilitation**, v. 67, n. 6, p. 387–389, jun. 1986.

MOKKINK, L. B.; BOERS, M.; VAN DER VLEUTEN, C. P. M.; BOUTER, L. M.; ALONSO, J.; PATRICK, D. L.; DE VET, H. C. W.; TERWEE, C. B. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. **BMC medical research methodology**, v. 20, n. 1, p. 293, 3 dez. 2020. https://doi.org/10.1186/s12874-020-01179-5.

MOKKINK, L. B.; DE VET, H. C. W.; PRINSEN, C. a. C.; PATRICK, D. L.; ALONSO, J.; BOUTER, L. M.; TERWEE, C. B. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. **Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation**, v. 27, n. 5, p. 1171–1179, maio 2018. https://doi.org/10.1007/s11136-017-1765-4.

MOKKINK, L. B.; TERWEE, C. B.; KNOL, D. L.; STRATFORD, P. W.; ALONSO, J.; PATRICK, D. L.; BOUTER, L. M.; DE VET, H. C. W. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. **BMC medical research methodology**, v. 6, p. 2, 24 jan. 2006. https://doi.org/10.1186/1471-2288-6-2.

MOKKINK, Lidwine B.; TERWEE, C. B.; PATRICK, D. L.; ALONSO, J.; STRATFORD, P. W.; KNOL, D. L.; BOUTER, L. M.; DE VET, H. C. W. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. **Quality of Life Research**, v. 19, n. 4, p. 539–549, 2010a. https://doi.org/10.1007/s11136-010-9606-8.

MOKKINK, Lidwine B.; TERWEE, C. B.; PATRICK, D. L.; ALONSO, J.; STRATFORD, P. W.; KNOL, D. L.; BOUTER, L. M.; DE VET, H. C. W. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. **Journal of Clinical Epidemiology**, v. 63, n. 7, p. 737–745, jul. 2010b. https://doi.org/10.1016/j.jclinepi.2010.02.006.

MULTIDISCIPLINARY STUDIES OF ILLNESS IN AGED PERSONS. II. A NEW CLASSIFICATION OF FUNCTIONAL STATUS IN ACTIVITIES OF DAILY LIVING. **Journal of Chronic Diseases**, v. 9, n. 1, p. 55–62, jan. 1959. https://doi.org/10.1016/0021-9681(59)90137-7.

MYERS, A. M.; HOLLIDAY, P. J.; HARVEY, K. A.; HUTCHINSON, K. S. Functional performance measures: are they superior to self-assessments? **Journal of Gerontology**, v. 48, n. 5, p. M196-206, set. 1993. https://doi.org/10.1093/geronj/48.5.m196.

NEPAL, G. M.; BASAULA, M.; SHARMA, S. Inter-rater reliability of Timed Up and Go test in older adults measured by physiotherapist and caregivers. **European Journal of Physiotherapy**, v. 22, n. 6, p. 325–331, 1 nov. 2020. https://doi.org/10.1080/21679169.2019.1623313.

OSTIR, G. V.; VOLPATO, S.; FRIED, L. P.; CHAVES, P.; GURALNIK, J. M.; WOMEN'S HEALTH AND AGING STUDY. Reliability and sensitivity to change assessed for a summary measure of lower body function: results from the Women's Health and Aging Study. **Journal of Clinical Epidemiology**, v. 55, n. 9, p. 916–921, set. 2002. https://doi.org/10.1016/s0895-4356(02)00436-5.

OUZZANI, M.; HAMMADY, H.; FEDOROWICZ, Z.; ELMAGARMID, A. Rayyan-a web and mobile app for systematic reviews. **Systematic Reviews**, v. 5, n. 1, p. 210, 5 dez. 2016. https://doi.org/10.1186/s13643-016-0384-4.

ÖZDEN, F.; ÖZKESKIN, M.; BAKIRHAN, S.; ŞAHIN, S. The test–retest reliability and concurrent validity of the 3-m backward walk test and 50-ft walk test in community-dwelling older adults. **Irish Journal of Medical Science (1971 -)**, v. 191, n. 2, p. 921–928, abr. 2022. https://doi.org/10.1007/s11845-021-02596-1.

PEEL, N. M.; KUYS, S. S.; KLEIN, K. Gait speed as a measure in geriatric assessment in clinical settings: a systematic review. **The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences**, v. 68, n. 1, p. 39–46, jan. 2013. https://doi.org/10.1093/gerona/gls174.

PODSIADLO, D.; RICHARDSON, S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. **Journal of the American Geriatrics Society**, v. 39, n. 2, p. 142–148, fev. 1991. https://doi.org/10.1111/j.1532-5415.1991.tb01616.x.

POLIT, D. F. w. **International Journal of Nursing Studies**, v. 52, n. 11, p. 1746–1753, nov. 2015. https://doi.org/10.1016/j.ijnurstu.2015.07.002.

PRINSEN, C. a. C.; MOKKINK, L. B.; BOUTER, L. M.; ALONSO, J.; PATRICK, D. L.; DE VET, H. C. W.; TERWEE, C. B. COSMIN guideline for systematic reviews of patient-reported outcome measures. **Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation**, v. 27, n. 5, p. 1147–1157, maio 2018. https://doi.org/10.1007/s11136-018-1798-3.

PRINSEN, C. A. C.; VOHRA, S.; ROSE, M. R.; BOERS, M.; TUGWELL, P.; CLARKE, M.; WILLIAMSON, P. R.; TERWEE, C. B. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. **Trials**, v. 17, n. 1, p. 449, 13 set. 2016. https://doi.org/10.1186/s13063-016-1555-2.

PUA, Y.-H.; MATCHAR, D. B. Physical Performance Predictor Measures in Older Adults With Falls-Related Emergency Department Visits. **Journal of the American Medical Directors Association**, v. 20, n. 6, p. 780–784, jun. 2019. https://doi.org/10.1016/j.jamda.2018.12.005.

REUBEN, D. B.; SIU, A. L. An objective measure of physical function of elderly outpatients. The Physical Performance Test. **Journal of the American Geriatrics Society**, v. 38, n. 10, p. 1105–1112, out. 1990. https://doi.org/10.1111/j.1532-5415.1990.tb01373.x.

REUBEN, D. B.; VALLE, L. A.; HAYS, R. D.; SIU, A. L. Measuring physical function in community-dwelling older persons: a comparison of self-administered, interviewer-administered, and performance-based measures. **Journal of the American Geriatrics Society**, v. 43, n. 1, p. 17–23, jan. 1995. https://doi.org/10.1111/j.1532-5415.1995.tb06236.x.

ROBERTSON, S.; KREMER, P.; AISBETT, B.; TRAN, J.; CERIN, E. Consensus on measurement properties and feasibility of performance tests for the exercise and sport sciences: a Delphi study. **Sports Medicine - Open**, v. 3, n. 1, p. 2, dez. 2017. https://doi.org/10.1186/s40798-016-0071-y.

ROZZINI, R.; FRISONI, G. B.; BIANCHETTI, A.; ZANETTI, O.; TRABUCCHI, M. Physical Performance Test and Activities of Daily Living scales in the assessment of health status in elderly people. **Journal of the American Geriatrics Society**, v. 41, n. 10, p. 1109–1113, out. 1993. https://doi.org/10.1111/j.1532-5415.1993.tb06460.x.

SAITO, Y.; NAKAMURA, S.; TANAKA, A.; WATANABE, R.; NARIMATSU, H.; CHUNG, U. Evaluation of the validity and reliability of the 10-meter walk test using a smartphone application among Japanese older adults. **Frontiers in Sports and Active Living**, v. 4, p. 904924, 4 out. 2022. https://doi.org/10.3389/fspor.2022.904924.

SAYERS, S. P.; GURALNIK, J. M.; NEWMAN, A. B.; BRACH, J. S.; FIELDING, R. A. Concordance and discordance between two measures of lower extremity function: 400 meter self-paced walk and SPPB. **Aging Clinical and Experimental Research**, v. 18, n. 2, p. 100–106, abr. 2006. https://doi.org/10.1007/BF03327424.

SCHOENE, D.; WU, S. M.-S.; MIKOLAIZAK, A. S.; MENANT, J. C.; SMITH, S. T.; DELBAERE, K.; LORD, S. R. Discriminative ability and predictive validity of the timed up and go test in identifying older people who fall: systematic review and meta-analysis. **Journal of the American Geriatrics Society**, v. 61, n. 2, p. 202–208, fev. 2013. https://doi.org/10.1111/jgs.12106.

SCHULER, P. B.; MARZILLI, T. S. Use of self-reports of physical fitness as substitutes for performance-based measures of physical fitness in older adults. **Perceptual and Motor Skills**, v. 96, n. 2, p. 414–420, abr. 2003. https://doi.org/10.2466/pms.2003.96.2.414.

SHERMAN, S. E.; REUBEN, D. Measures of Functional Status in Community-Dwelling Elders. **Journal of General Internal Medicine**, v. 13, n. 12, p. 817–823, dez. 1998. https://doi.org/10.1046/j.1525-1497.1998.00245.x.

SHERWOOD, J. J.; INOUYE, C.; WEBB, S. L.; O, J. Reliability and Validity of the Sit-to-Stand as a Muscular Power Measure in Older Adults. **Journal of Aging and Physical Activity**, v. 28, n. 3, p. 455–466, 1 jun. 2020. https://doi.org/10.1123/japa.2019-0133.

SIMONSICK, E. M.; NEWMAN, A. B.; NEVITT, M. C.; KRITCHEVSKY, S. B.; FERRUCCI, L.; GURALNIK, J. M.; HARRIS, T.; HEALTH ABC STUDY GROUP. Measuring higher level physical function in well-functioning older adults: expanding familiar approaches in the Health ABC study. **The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences**, v. 56, n. 10, p. M644-649, out. 2001. https://doi.org/10.1093/gerona/56.10.m644.

SUCHY, Y.; KRAYBILL, M. L.; FRANCHOW, E. Instrumental activities of daily living among community-dwelling older adults: discrepancies between self-report and performance are mediated by cognitive reserve. **Journal of Clinical and Experimental Neuropsychology**, v. 33, n. 1, p. 92–100, jan. 2011. https://doi.org/10.1080/13803395.2010.493148.

SUCHY, Y.; WILLIAMS, P. G.; KRAYBILL, M. L.; FRANCHOW, E.; BUTNER, J. Instrumental activities of daily living among community-dwelling older adults: personality associations with self-report, performance, and awareness of functional difficulties. **The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences**, v. 65, n. 5, p. 542–550, set. 2010. https://doi.org/10.1093/geronb/gbq037.

TERWEE, C. B.; PRINSEN, C. a. C.; CHIAROTTO, A.; WESTERMAN, M. J.; PATRICK, D. L.; ALONSO, J.; BOUTER, L. M.; DE VET, H. C. W.; MOKKINK, L. B. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. **Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation**, v. 27, n. 5, p. 1159–1170, maio 2018. https://doi.org/10.1007/s11136-018-1829-0.

TERWEE, C. B.; BOT, S. D. M.; DE BOER, M. R.; VAN DER WINDT, D. A. W. M.; KNOL, D. L.; DEKKER, J.; BOUTER, L. M.; DE VET, H. C. W. Quality criteria were proposed for measurement properties of health status questionnaires. **Journal of Clinical Epidemiology**, v. 60, n. 1, p. 34–42, jan. 2007. https://doi.org/10.1016/j.jclinepi.2006.03.012.

TERWEE, C. B.; JANSMA, E. P.; RIPHAGEN, I. I.; DE VET, H. C. W. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. **Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation**, v. 18, n. 8, p. 1115–1123, out. 2009. https://doi.org/10.1007/s11136-009-9528-5.

TERWEE, C. B.; PEIPERT, J. D.; CHAPMAN, R.; LAI, J.-S.; TERLUIN, B.; CELLA, D.; GRIFFITHS, P.; MOKKINK, L. B. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. **Quality of Life Research**, v. 30, n. 10, p. 2729–2754, 2021. https://doi.org/10.1007/s11136-021-02925-y.

WINOGRAD, C. H.; LEMSKY, C. M.; NEVITT, M. C.; NORDSTROM, T. M.; STEWART, A. L.; MILLER, C. J.; BLOCH, D. A. Development of a physical performance and mobility examination. **Journal of the American Geriatrics Society**, v. 42, n. 7, p. 743–749, jul. 1994. https://doi.org/10.1111/j.1532-5415.1994.tb06535.x.

WORLD HEALTH ORGANIZATION. **World report on ageing and health**. [*S. l.*]: World Health Organization, 2015. Disponível em: https://apps.who.int/iris/handle/10665/186463. Acesso em: 21 set. 2022.

ZANKER, J.; SCOTT, D.; ALAJLOUNI, D.; KIRK, B.; BIRD, S.; DEBRUIN, D.; VOGRIN, S.; BLIUC, D.; TRAN, T.; CAWTHON, P.; DUQUE, G.; CENTER, J. R. Mortality, falls and slow walking speed are predicted by different muscle strength and physical performance measures in women and men. **Archives of Gerontology and Geriatrics**, v. 114, p. 105084, nov. 2023. https://doi.org/10.1016/j.archger.2023.105084.

ZANKER, J.; SIM, M.; ANDERSON, K.; BALOGUN, S.; BRENNAN-OLSEN, S. L.; DENT, E.; DUQUE, G.; GIRGIS, C. M.; GROSSMANN, M.; HAYES, A.; HENWOOD, T.; HIRANI, V.; INDERJEETH, C.; IULIANO, S.; KEOGH, J.; LEWIS, J. R.; LYNCH, G. S.; PASCO, J. A.; PHU, S.; … SCOTT, D. Consensus guidelines for sarcopenia prevention, diagnosis and management in Australia and New Zealand. **Journal of Cachexia, Sarcopenia and Muscle**, v. 14, n. 1, p. 142–156, 9 nov. 2022. https://doi.org/10.1002/jcsm.13115.

8 APÊNDICE A: ESTRATÉGIA DE BUSCA NA BASE DE DADO PUBMED

((("Aged"[MeSH Terms] OR "aged, 80 and over"[MeSH Terms] OR "Frail Elderly"[MeSH Terms] OR "Aging"[MeSH Terms] OR ("elderly frail"[All Fields] OR "Frail Elders"[All Fields] OR ("Frail Elderly"[MeSH Terms] OR ("frail"[All Fields] AND "elderly"[All Fields]) OR "Frail Elderly"[All Fields] OR ("elder"[All Fields] AND "frail"[All Fields])) OR ("Frail Elderly"[MeSH Terms] OR ("frail"[All Fields] AND "elderly"[All Fields]) OR "Frail Elderly"[All Fields] OR ("elders"[All Fields] AND "frail"[All Fields])) OR "Frail Elder"[All Fields] OR "functionally impaired elderly"[All Fields] OR "elderly functionally impaired"[All Fields] OR "functionally impaired elderly"[All Fields] OR "Frail Older Adults"[All Fields] OR ("Frail Elderly"[MeSH Terms] OR ("frail"[All Fields] AND "elderly"[All Fields]) OR "Frail Elderly"[All Fields] OR ("adult"[All Fields] AND "frail"[All Fields] AND "older"[All Fields])) OR "adults frail older"[All Fields] OR "Frail Older Adult"[All Fields] OR "older adult frail"[All Fields] OR "older adults frail"[All Fields] OR "Oldest Old"[All Fields] OR "Older People"[All Fields] OR "Elderly People"[All Fields] OR "very elderly"[All Fields] OR "Old People"[All Fields] OR ("geriatric"[All Fields] OR "geriatrics"[MeSH Terms] OR "geriatrics"[All Fields]) OR ("senior"[All Fields] OR "seniorities"[All Fields] OR "seniority"[All Fields] OR "seniors"[All Fields]) OR "Aging"[All Fields] OR "ageing"[All Fields])) AND ("psychometrical"[All Fields] OR "psychometrically"[All Fields] OR "psychometrics"[MeSH Terms] OR "psychometrics"[All Fields] OR "psychometric"[All Fields] OR ("measurability"[All Fields] OR "measurable"[All Fields] OR "measurably"[All Fields] OR "measure s"[All Fields] OR "measureable"[All Fields] OR "measured"[All Fields] OR "measurement"[All Fields] OR "measurement s"[All Fields] OR "measurements"[All Fields] OR "measurer"[All Fields] OR "measurers"[All Fields] OR "measuring"[All Fields] OR "measurings"[All Fields] OR "measurment"[All Fields] OR "measurments"[All Fields] OR "weights and measures"[MeSH Terms] OR ("weights"[All Fields] AND "measures"[All Fields]) OR "weights and measures"[All Fields] OR "measure"[All Fields] OR "measures"[All Fields]) OR ("reliabilities"[All Fields] OR "reliability"[All Fields] OR "reliable"[All Fields] OR "reliablity"[All Fields] OR "reliably"[All Fields]) OR ("repeatabilities"[All Fields] OR "repeatability"[All Fields] OR "repeatable"[All Fields] OR "repeated"[All Fields] OR "repeatibility"[All Fields]) OR ("reproducability"[All Fields] OR "reproducibilities"[All Fields] OR "reproducibility"[All Fields] OR "reproducible"[All Fields]) OR "measurement error"[All Fields] OR ("consistence"[All Fields] OR "consistences"[All Fields] OR "consistencies"[All Fields] OR "consistency"[All Fields]

OR "consistent"[All Fields] OR "consistently"[All Fields]) OR "smallest worthwhile change"[All Fields] OR "minimal detectable change"[All Fields] OR "typical error"[All Fields] OR ("useful"[All Fields] OR "usefulness"[All Fields]) OR ("valid"[All Fields] OR "validate"[All Fields] OR "validated"[All Fields] OR "validates"[All Fields] OR "validating"[All Fields] OR "validation"[All Fields] OR "validational"[All Fields] OR "validations"[All Fields] OR "validator"[All Fields] OR "validators"[All Fields] OR "validities"[All Fields] OR "validity"[All Fields]) OR ("logic"[MeSH Terms] OR "logic"[All Fields] OR "logics"[All Fields] OR "logical"[All Fields] OR "logically"[All Fields]) OR ("construct s"[All Fields] OR "constructed"[All Fields] OR "constructing"[All Fields] OR "construction"[All Fields] OR "constructions"[All Fields] OR "constructive"[All Fields] OR "constructively"[All Fields] OR "constructs"[All Fields] OR "construct"[All Fields]) OR ("converge"[All Fields] OR "converged"[All Fields] OR "convergence"[All Fields] OR "convergences"[All Fields] OR "convergencies"[All Fields] OR "convergency"[All Fields] OR "convergent"[All Fields] OR "convergently"[All Fields] OR "convergents"[All Fields] OR "converges"[All Fields] OR "converging"[All Fields]) OR ("discriminabilities"[All Fields] OR "discriminability"[All Fields] OR "discriminable"[All Fields] OR "discriminably"[All Fields] OR "discriminance"[All Fields] OR "discriminant"[All Fields] OR "discriminants"[All Fields] OR "discriminate"[All Fields] OR "discriminated"[All Fields] OR "discriminates"[All Fields] OR "discriminating"[All Fields] OR "discrimination, psychological"[MeSH Terms] OR ("discrimination"[All Fields] AND "psychological"[All Fields]) OR "psychological discrimination"[All Fields] OR "discrimination"[All Fields] OR "discriminations"[All Fields] OR "discriminative"[All Fields] OR "discriminatively"[All Fields] OR "discriminator"[All Fields] OR "discriminators"[All Fields]) OR "gold standard"[All Fields] OR ("level"[All Fields] OR "levels"[All Fields]) OR ("reference standards"[MeSH Terms] OR ("reference"[All Fields] AND "standards"[All Fields]) OR "reference standards"[All Fields] OR "standardization"[All Fields] OR "standard"[All Fields] OR "standard s"[All Fields] OR "standardisation"[All Fields] OR "standardisations"[All Fields] OR "standardise"[All Fields] OR "standardised"[All Fields] OR "standardises"[All Fields] OR "standardising"[All Fields] OR "standardization s"[All Fields] OR "standardizations"[All Fields] OR "standardize"[All Fields] OR "standardized"[All Fields] OR "standardizes"[All Fields] OR "standardizing"[All Fields] OR "standards"[MeSH Subheading] OR "standards"[All Fields]) OR "physical performance"[All Fields] OR "functional performance"[All Fields]) AND ("Timed Up and Go test"[All Fields] OR "Timed Up Go test"[All Fields] OR "Get Up and Go test"[All Fields] OR "TUG"[All Fields] OR "GUG"[All Fields] OR "TGUG"[All Fields] OR "TGUGT"[All Fields]

OR "TUGT"[All Fields] OR "modified TUG"[All Fields] OR "ITUG"[All Fields] OR "sit-to-stand TEST"[All Fields] OR "Sit-tostand"[All Fields] OR "stand-to-sit"[All Fields] OR "chair stand"[All Fields] OR "STS"[All Fields] OR "5STS"[All Fields] OR "30STS"[All Fields] OR "five-repetition sit-to-stand test"[All Fields] OR "5 times sit-to-stand test"[All Fields] OR "gait speed"[All Fields] OR "Walking Speed"[MeSH Terms] OR "Walking Speed"[All Fields] OR "Gait Velocity Test"[All Fields] OR "10 meter walking test"[All Fields] OR "6MWT"[All Fields] OR "10MWT"[All Fields] OR "6 meter walking test"[All Fields] OR "5 meter walking test"[All Fields] OR "4 meter walking test"[All Fields] OR "gait velocity"[All Fields] OR "vertical jump test"[All Fields] OR "power test"[All Fields] OR ("countermovement"[All Fields] OR "countermovements"[All Fields]) OR "countermovement jump test"[All Fields] OR "squat jump"[All Fields] OR "squat jump test"[All Fields] OR "agility test"[All Fields] OR "change of Direction"[All Fields] OR "change of direction test"[All Fields]) AND "intraclass correlation coefficient"[All Fields]) NOT ("pulmonary disease, chronic obstructive"[MeSH Terms] OR "pulmonary disease chronic obstructive"[All Fields] OR "Diabetes Mellitus"[MeSH Terms] OR "hydrocephalus, normal pressure"[MeSH Terms] OR "arthroplasty, replacement, hip"[MeSH Terms] OR "arthroplasty, replacement, knee"[MeSH Terms] OR "Diabetes Mellitus"[All Fields] OR "hydrocephalus normal pressure"[All Fields] OR "arthroplasty replacement hip"[All Fields] OR "arthroplasty replacement knee"[All Fields] OR "Stroke"[MeSH Terms] OR "HIV"[MeSH Terms] OR "Neoplasms"[MeSH Terms] OR "Stroke"[All Fields] OR "HIV"[All Fields] OR "Neoplasms"[All Fields] OR "Parkinson Disease"[All Fields] OR "Dementia"[All Fields] OR "Alzheimer Disease"[All Fields] OR "Diabetic Neuropathies"[MeSH Terms] OR "Diabetic Neuropathies"[All Fields] OR "osteoarthritis"[MeSH Terms] OR "osteoarthritis"[All Fields] OR "osteoarthrosis"[All Fields] OR "knee osteoarthritis"[All Fields] OR "osteoarthritis"[MeSH Terms] OR "osteoarthritis"[All Fields] OR "osteoarthritides"[All Fields] OR "hip osteoarthrosis"[All Fields] OR "interstitial lung disease"[All Fields] OR ("child"[MeSH Terms] OR "child"[All Fields] OR "children"[All Fields] OR "child s"[All Fields] OR "children s"[All Fields] OR "childrens"[All Fields] OR "childs"[All Fields] OR "spinal muscular atrophy"[All Fields]) OR "chronic kidney disease"[All Fields])) NOT "systematic review"[Publication Type]

9 APÊNDICE B: LISTA DE VERIFICAÇÃO PRISMA (2020)

Seção e Tópico	Item#	Item da lista de verificação	Local onde o item é relatado
TÍTULO			
Título	1	Identifique o relatório como uma revisão sistemática.	Pag. 1, 2 e 3.
RESUMO			
Resumo	2	Consulte a lista de verificação PRISMA 2020 para resumos.	Pag. 7 e 8
INTRODUÇÃO			
Justificativa	3	Descreva a justificativa para a revisão no contexto do conhecimento existente.	Pag. 18
Objetivos	4	Forneça uma declaração explícita do (s) objetivo (s) ou da (s) questão (ões) que a revisão aborda.	Pag. 19
MÉTODOS			
Critério de eleição	5	Especifique os critérios de inclusão e exclusão para a revisão e como os estudos foram agrupados para as sínteses.	Pag. 31, 1º-2º parág.
Fontes de informação	6	Especifique todas as bases de dados, registros, sites, organizações, listas de referência e outras fontes pesquisadas ou consultadas para identificar estudos. Especifique a data em que cada fonte foi pesquisada ou consultada pela última vez.	Pag. 29, 2º parág.
Procurar estratégia	7	Apresente as estratégias de pesquisa completas para todos os bancos de dados, registros e sites, incluindo quaisquer filtros e limites usados.	Pag. 29, 3º e 4º parág; Pag. 30.
Processo de seleção	8	Especifique os métodos usados para decidir se um estudo atendeu aos critérios de inclusão da revisão, incluindo quantos revisores selecionaram cada registro e cada relatório recuperado, se trabalharam de forma independente e, se aplicável, detalhes das ferramentas de automação usadas no processo.	Pág. 31, 3º parág.
Processo de coleta de dados	9	Especifique os métodos usados para coletar dados de relatórios, incluindo quantos revisores coletaram dados de cada relatório, se eles trabalharam de forma independente, quaisquer processos para obter ou confirmar dados dos investigadores do estudo e, se aplicável, detalhes das ferramentas de automação usadas no processo.	Pág. 31, 3º parág; Pag. 32, 1º parág.
Itens de dados	10a	Liste e defina todos os resultados para os quais os dados foram buscados. Especifique se todos os resultados que eram compatíveis com cada domínio de resultado em cada estudo foram buscados (por exemplo, para todas as medidas, pontos de tempo, análises) e, se não, os métodos usados para decidir quais resultados coletar.	Pág. 31, 3º parág; Pag. 32, 1º parág.
	10b	Liste e defina todas as outras variáveis para as quais os dados foram buscados (por exemplo, características do participante e da intervenção, fontes de financiamento). Descreva quaisquer suposições feitas sobre qualquer informação ausente ou pouco clara.	Pág. 31, 3º parág; Pag. 32, 1º parág.
Estudo de risco	11	Especifique os métodos usados para avaliar o risco de viés nos estudos incluídos, incluindo detalhes da (s) ferramenta (s) usada (s), quantos	Pag. 32, 2º

Seção e Tópico	Item#	Item da lista de verificação	Local onde o item é relatado
de avaliação de viés		revisores avaliaram cada estudo e se trabalharam de forma independente e, se aplicável, detalhes das ferramentas de automação usadas no processo.	e 3º parág; Pag. 33; Pag. 34.
Medidas de efeito	12	Especifique para cada resultado a (s) medida (s) de efeito (por exemplo, razão de risco, diferença média) usada na síntese ou apresentação dos resultados.	Pag. 33
Métodos de síntese	13a	Descreva os processos usados para decidir quais estudos eram elegíveis para cada síntese (por exemplo, tabulando as características da intervenção do estudo e comparando com os grupos planejados para cada síntese (item # 5)).	n/a
	13b	Descreva quaisquer métodos necessários para preparar os dados para apresentação ou síntese, como tratamento de estatísticas de resumo ausentes ou conversões de dados.	n/a
	13c	Descreva quaisquer métodos usados para tabular ou exibir visualmente os resultados de estudos e sínteses individuais.	n/a
	13d	Descreva quaisquer métodos usados para sintetizar os resultados e forneça uma justificativa para-a (s) escolha (ões). Se uma meta-análise foi realizada, descreva o (s) modelo (s), método (s) para identificar a presença e extensão da heterogeneidade estatística e o (s) pacote (s) de software usado (s).	n/a
	13e	Descreva quaisquer métodos usados para explorar as possíveis causas da heterogeneidade entre os resultados do estudo (por exemplo, análise de subgrupo, meta-regressão).	n/a
	13f	Descreva quaisquer análises de sensibilidade conduzidas para avaliar a robustez dos resultados sintetizados.	n/a
Avaliação de viés de relatório	14	Descreva quaisquer métodos usados para avaliar o risco de viés devido à falta de resultados em uma síntese (decorrente de vieses de relatórios).	n/a
Avaliação de certeza	15	Descreva quaisquer métodos usados para avaliar a certeza (ou confiança) no corpo de evidências para um resultado.	Pag.34
RESULTADOS			
Seleção de estudos	16a	Descreva os resultados do processo de busca e seleção, desde o número de registros identificados na busca até o número de estudos incluídos na revisão, de preferência por meio de um fluxograma.	Pag. 35
	16b	Cite estudos que possam parecer atender aos critérios de inclusão, mas que foram excluídos, e explique por que foram excluídos.	Pag. 35
Características do estudo	17	Cite cada estudo incluído e apresente suas características.	Pag. 36-56
Risco de viés em estudos	18	Apresentar avaliações de risco de viés para cada estudo incluído.	Pag. 57-69
Resultados de estudos individuais	19	Para todos os resultados, apresente, para cada estudo: (a) estatísticas resumidas para cada grupo (quando apropriado) e (b) uma estimativa de efeito e sua precisão (por exemplo, intervalo de confiança / credibilidade), idealmente usando tabelas ou gráficos estruturados.	n/a
Resultados de sínteses	20a	Para cada síntese, resuma brevemente as características e o risco de viés entre os estudos contribuintes.	Pag. 59-60
	20b	Apresentar os resultados de todas as sínteses estatísticas realizadas. Se a meta-análise foi feita, apresente para cada uma a estimativa	n/a

Seção e Tópico	Item#	Item da lista de verificação	Local onde o item é relatado
		resumida e sua precisão (por exemplo, intervalo de confiança / credibilidade) e medidas de heterogeneidade estatística. Se estiver comparando grupos, descreva a direção do efeito.	
	20c	Apresentar os resultados de todas as investigações das possíveis causas de heterogeneidade entre os resultados do estudo.	n/a
	20d	Apresentar os resultados de todas as análises de sensibilidade conduzidas para avaliar a robustez dos resultados sintetizados.	n/a
Polarização de relatórios	21	Apresente avaliações de risco de viés devido a resultados ausentes (decorrentes de vieses de relatórios) para cada síntese avaliada.	n/a
Certeza de evidência	22	Apresentar avaliações de certeza (ou confiança) no corpo de evidências para cada resultado avaliado.	Pag. 70-75
DISCUSSÃO			
Discussão	23a	Forneça uma interpretação geral dos resultados no contexto de outras evidências.	Pag.75, 1º parág.
	23b	Discuta quaisquer limitações das evidências incluídas na revisão.	Pag.75, 2º parág; Pag. 76, 1º e 2º parág.
	23c	Discuta quaisquer limitações dos processos de revisão usados.	Pag. 76, 4º parág; Pag. 77, 1º e 2º parág.
	23d	Discuta as implicações dos resultados para a prática, política e pesquisas futuras.	Pag. 77, 3º parág.
OUTRA INFORMAÇÃO			
Registro e protocolo	24a	Forneça informações de registro para a revisão, incluindo nome de registro e número de registro, ou declare que a revisão não foi registrada.	Pag. 28, 1º parág.
	24b	Indique onde o protocolo de revisão pode ser acessado ou indique que um protocolo não foi preparado.	Pag. 28, 1º parág.
	24c	Descreva e explique quaisquer alterações nas informações fornecidas no registro ou no protocolo.	n/a
Apoio, suporte	25	Descreva as fontes de apoio financeiro ou não financeiro para a revisão e a função dos financiadores ou patrocinadores na revisão.	n/a
Interesses competitivos	26	Declare quaisquer interesses conflitantes dos autores da revisão.	n/a
Disponibilidade de dados, código e outros materiais	27	Relate quais dos seguintes itens estão disponíveis publicamente e onde podem ser encontrados: modelos de formulários de coleta de dados; dados extraídos dos estudos incluídos; dados usados para todas as análises; código analítico; quaisquer outros materiais usados na revisão.	n/a

A partir de: Página MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. A declaração PRISMA 2020: uma diretriz atualizada para relatar revisões sistemáticas. BMJ 2021; 372: n71. doi: 10.1136 / bmj.n71

10 APENDICE C: AVALIAÇÃO DOS ESTUDOS LIDOS NA ÍNTEGRA

Acessar pelo link: PLANILHA_LEITURA DOS ESTUDOS NA ÍNTEGRA_AV.1_AV.2.xlsx

11 APENDICE D: PLANILHA DE EXTRAÇÃO DE DADOS DOS ESTUDOS

Acessar pelo link: PLANILHA_EXTRAÇÃO DOS DADOS_Av.1_Av.2.xlsx

12 APENDICE E: AVALIAÇÃO DO COSMIN RISK OF BIAS CHECKLIST

Caixas 6 a 9 acessar pelo link: PLANILHA COSMIN RISK OF BIAS CHECKLIST Av.1 Av.2.xlsx

Caixa 6 - Confiabilidade

Caixa 7 - Erro de Medição

Caixa 8 - Validade de Critério

Caixa 9 - Validade de Construto

13 APENDICE F: AVALIAÇÃO INDIVIDUAL E SUMARIZADA DOS CRITÉRIOS PARA BOAS PROPRIEDADES DE MEDIÇÃO E METODOLOGIA GRADE

Acessar pelo link: PLANILHA_AVALIAÇÃO INDIVIDUAL_SUMARIZADA_GRADE_AV.1_AV.2.xlsx

14 APÊNDICE G: LIVRO AVALIAÇÃO FÍSICO FUNCIONAL EM IDOSOS

Para acessar clique aqui