

Teste de sentar-levantar: estudos de fidedignidade

Sitting-rising test: reliability studies

Vitor Agnew Lira ¹

Claudio Gil Soares de Araújo ^{1,2}

Resumo:

[1] Lira, V. A. e Araújo, C.G.S. Teste de sentar-levantar: estudos de fidedignidade, Rev. Bras. Ciên. e Mov. 8 (2): 11-20, 2000.

Há pouca documentação sobre a influência de um estilo de vida sedentário na habilidade em realizar atividades simples e rotineiras. O objetivo deste trabalho foi investigar as fidedignidades intra e interavaliador e intra e interdias do Teste de Sentar-Levantar (TSL). O TSL avalia a destreza nas ações de sentar e levantar do solo, independentemente, utilizando uma escala ordinal de 0 a 5. Para cada apoio utilizado nas ações, um ponto é perdido e, caso haja desequilíbrio, mais meio ponto é subtraído do escore máximo de 5. O melhor desempenho, em duas tentativas, representa o escore final para cada ação. Primeiramente, a fidedignidade interavaliadores foi abordada através de dois avaliadores, graduando simultaneamente os desempenhos de 29 indivíduos. Depois, três avaliadores assistiram a uma fita de vídeo, contendo 10 execuções para cada uma das ações, com todos os escores simulados seguindo uma ordem randômica estratificada. Após 10 meses, um dos avaliadores assistiu à fita novamente, a fim de estabelecer a fidedignidade intravaliador. Na determinação da fidedignidade interdias, 10 indivíduos, aparentemente saudáveis, foram avaliados em quatro dias distintos e próximos. Em um dia aleatório para cada indivíduo, 10 avaliações consecutivas foram conduzidas. Nenhuma das abordagens revelou diferenças significativas entre avaliadores ou avaliações, para ambas as ações ($p > 0,05$). Verificou-se, também, ocorrências próximas ao modelo teórico esperado de 80%, 10%, 7% e 3%, respectivamente, para concordância absoluta ou manutenção dos escores, discordância quanto à presença de desequilíbrio, discordância quanto ao uso de apenas um apoio e discordância importante ou demais divergências ($p > 0,11$). Apenas a ação de levantar, no mesmo dia, apresentou maior homogeneidade na distribuição entre concordância e níveis de discordância ($p < 0,01$), talvez por razão motivacional. No estudo da fidedignidade intra e interdias, não foram encontradas diferenças entre a frequência total dos escores e a frequência dos mesmos, em cada avaliação ($p > 0,34$). Assim, o melhor escore para cada in-

divíduo ocorria aleatoriamente, entre dias ou avaliações. Conclui-se que o TSL é fidedigno e, portanto, reproduzível. Parece não ser possível a melhoria na eficiência de execução das ações, em curto intervalo de tempo, o que reforça a confiabilidade e a estabilidade dos resultados. Estudos futuros devem abordar a fidedignidade do teste em grupos populacionais distintos.

UNITERMOS: Teste de Sentar-Levantar, fidedignidade, aptidão física, avaliação funcional.

Abstract

[2] Lira, V. A. and Araújo, C.G.S. Sitting-rising test: reliability studies, Rev. Bras. Ciên. e Mov. 8 (2): 11-20, 2000.

The literature presents poor documentation about the influence of a sedentary lifestyle on the ability in performing simple and routine tasks. The objective of this study is to investigate the intrarater, interrater, as well as trial-to-trial and day-to-day reliabilities of the Sitting-Rising Test (SRT). The SRT evaluates the ability in sitting and rising from the floor independently, using an ordinal scale ranging from 0 to 5. For each extra support used in the actions one point is lost, and if there is any unbalance, a further .5 is withdrawn from the maximal score 5. The best performance in two trials represents the final score for each action. Firstly, interrater reliability was approached through two evaluators scoring simultaneously the executions of 29 individuals. Then, three evaluators watch a video-taped film containing 10 executions for each action, with every possible score simulated, obeying a stratified random. After 10 months, one of the evaluators watched the film again in order to establish the intrarater reliability. Determining trial-to-trial and day-to-day reliabilities, 10 apparently healthy individuals were evaluated in four distinct but close days. At one random day for each individual, 10 consecutive evaluations were also conducted. None of the approaches revealed significant differences among evaluators, or evaluations, for both actions ($p > .05$). We also verified similar occurrences to the expected theoretical model of

¹ Programa de Pós-Graduação em Educação Física da Universidade Gama Filho

² Clínica de Medicina do Exercício (Clinimex)
Rio de Janeiro, RJ, Brasil

End: Clinimex
Rua Siqueira Campos, 93/101
22031-070, Rio de Janeiro – RJ, Brasil
E-mail: cgaraujo@iis.com.br

80%, 10%, 7% and 3%, respectively, for absolute agreement or maintenance of scores, disagreement concerning the presence of unbalance, disagreement concerning the use of just one support and important disagreement or further divergences ($p > .11$). Only the rising action in the same day presented a larger homogeneity in the distribution among agreement and levels of disagreement ($p < .01$), maybe due to motivational purposes. In the study of trial-to-trial and day-to-day reliabilities, no significant differences were found between the total frequency of scores and their frequency in each evaluation ($p > .34$). Thus, the best score for each individual happened randomly among days, or evaluations. We conclude that the SRT is reliable and therefore reproducible. A short time improvement in the efficiency of actions execution is unlikely, which underlines the reliability and stability of the scores. Future studies should approach the test reliability in different populations.

KEYWORDS: Sitting-Rising Test, reliability, physical fitness, functional evaluation.

Introdução

Historicamente, tem sido dada considerável atenção à influência deletéria de um estilo de vida sedentário e de uma baixa condição física no desenvolvimento de doenças crônico-degenerativas, tais como a doença arterial coronariana, a hipertensão arterial sistêmica, a obesidade e alguns tipos de câncer (6,17). Entretanto, a literatura é mais escassa no que se refere à análise desses efeitos sobre os movimentos e as atividades cotidianas, em pessoas saudáveis e não saudáveis (8,21).

A manutenção de uma condição muscular funcional mínima é vital para a qualidade de vida relacionada à saúde (18,19) e, em particular, para pessoas idosas (13) e populações especiais em que a massa corporal magra encontra-se reduzida, como na maioria dos indivíduos portadores do HIV (9).

Sentar e levantar estão dentre as atividades mais rotineiramente praticadas na vida diária, e o desempenho nessas ações apresenta uma relação estreita com o risco de queda (21) e, também, com a dificuldade de se levantar do solo logo após uma queda, por exemplo, não tendo ocorrido lesões importantes (1). Níveis mínimos de potência muscular, coordenação, equilíbrio (20) e flexibilidade (19) parecem ser necessários para o sucesso nessas ações, bem como em outras atividades cotidianas. Recentemente, objetivando avaliar o desempenho nas referidas ações, Araújo (3) propôs um método simples, denominado Teste de Sentar-Levantar (TSL).

O TSL envolve os atos de sentar e levantar do solo, comuns nos primeiros anos de vida, mas progressivamente menos presentes no cotidiano, de acordo com o passar dos anos. A lógica que permeia a avaliação é a de que, quanto maior a dificuldade do indivíduo em realizar os atos, mais apoios no solo e no próprio corpo serão utilizados. A aplicação do teste demanda cerca de um minuto, e a gradação dos atos é extremamente simples, uma vez que cada apoio

utilizado resulta na redução de um ponto da nota máxima e, havendo desequilíbrio perceptível, mais meio ponto é subtraído. Todas as características apontadas, até então, conferem ao TSL um perfil apropriado para testagem de rastreamento, situações nas quais muitos indivíduos devem ser avaliados em curto período de tempo, priorizando a identificação de resultados baixos ou insuficientes. Esses tipos de teste devem possuir elevada sensibilidade e, se possível, alta especificidade para a avaliação da aptidão física e/ou prontidão para a prática de exercícios, em várias populações. Essas informações devem ainda ser obtidas rápida e facilmente, de forma não invasiva, com baixo risco para o avaliado e a um custo reduzido (3).

A discussão sobre a aplicabilidade de um teste deve ser precedida, entretanto, da investigação sobre sua validade e fidedignidade. Como validade, entende-se a capacidade do procedimento medir, efetivamente, a variável ou conjunto de variáveis para as quais foi desenvolvido (5). Quanto à fidedignidade, significa fornecer resultados que sejam consistentes, quando obtidos nas mesmas condições ou em condições bem semelhantes (22). Atkinson e Nevill (4) apontam que a fidedignidade deve ser o primeiro fator a ser investigado em um novo instrumento de avaliação, uma vez que o mesmo será válido somente se apresentar uma consistência aceitável em seus resultados.

A fidedignidade talvez possa ser caracterizada de três formas principais: intradia, interdias e interavaliadores, que são determinadas a partir de estudos com características distintas, mas que acabam por fornecer informações complementares. A reprodutibilidade de medidas tomadas em um mesmo dia, com curto intervalo de tempo, porém suficiente para a recuperação, pode ser considerada como a fidedignidade intradia ou estabilidade de um teste. A variabilidade em medidas tomadas em dias diferentes e próximos consiste na fidedignidade interdias e também fornece parâmetros para o estabelecimento da estabilidade do procedimento. Já o grau de concordância entre dois ou mais avaliadores sobre as medidas é freqüentemente abordado como a fidedignidade interavaliadores ou objetividade de um teste (4). Como ressalva, é importante comentar que os dados sobre a consistência de medidas feitas pelo mesmo avaliador podem fornecer uma maior fundamentação para a fidedignidade interavaliadores. Assim, o presente trabalho foi constituído por quatro estudos independentes, voltados para a análise da fidedignidade dos resultados obtidos a partir do TSL.

Os objetivos do presente estudo foram investigar as fidedignidades: a) interavaliadores; b) intravaliador; c) intradia e; d) interdias.

Materiais e métodos

Procurando atender os objetivos mencionados, todas as investigações foram realizadas em indivíduos voluntários e assintomáticos para comprometimentos no sistema locomotor. Primeiramente, faremos a descrição metodológica do TSL e, posteriormente, serão abordados os protocolos relacionados a cada um dos quatro estudos.

Teste de Sentar-Levantar:

O Teste de Sentar-Levantar (TSL) é um procedimento simples, que tem como objetivo avaliar a destreza na execução das ações de sentar e levantar do solo. A avaliação é feita separadamente, para cada ação, atribuindo-se escores independentes.

O TSL deve ser administrado em uma superfície plana, não escorregadia. O avaliador deve posicionar-se à frente e em diagonal ao avaliado, procurando uma visão completa de seus movimentos e a fim de fornecer segurança ao mesmo. Este último aspecto é especialmente importante na avaliação de pessoas idosas ou de indivíduos que sabidamente tenham sofrido perda importante e recente de massa corporal magra (portadores de lesões de membros inferiores, indivíduos submetidos a grandes cirurgias etc).

O avaliado deve estar descalço e sem meias, trazendo roupas que não restrinjam o arco de movimento das articulações do tornozelo, joelho, quadril e tronco. Preferencialmente, um colchonete deve ser posicionado atrás do avaliado, visando a minimizar o impacto do quadril com o solo, durante a ação de sentar. Observa-se, contudo, que os pés do avaliado fiquem fora do colchonete, para evitar eventuais desequilíbrios pelo deslize desse sobre a superfície do solo.

FIGURA 1. Graduação do Teste de Sentar-Levantar (TSL)

SENTAR:	LEVANTAR:
5 - sem apoios	5 - sem apoios
4 - com 1 apoio	4 - com 1 apoio
3 - com 2 apoios	3 - com 2 apoios
2 - com 3 apoios	2 - com 3 apoios
1 - com 4 apoios	1 - com 4 apoios
0 - com mais de 4 apoios ou com ajuda externa	0 - com mais de 4 apoios ou com ajuda externa
Havendo desequilíbrio: subtrai-se 0,5 ponto -4,5; 3,5; 2,5; 1,5; 0,5.	Havendo desequilíbrio: subtrai-se 0,5 ponto -4,5; 3,5; 2,5; 1,5; 0,5.

Para a graduação da destreza independente nas duas ações, utiliza-se uma escala de mensuração ordinal, descontínua e crescente, de zero a cinco, com 10 intervalos de meio ponto (Tabela 1). A nota máxima cinco, corresponde à ação de sentar ou à de levantar, realizada sem a utilização de apoio extra (mão, braço e joelho) sem qualquer desequilíbrio corporal perceptível. Para cada apoio extra utilizado no solo e/ou no próprio corpo, para facilitar a execução ou evitar choque com o solo, subtrai-se um ponto da nota máxima. O desequilíbrio perceptível resulta em um decréscimo de mais meio ponto. Por exemplo, se um indivíduo senta no solo equilibradamente, utilizando apenas um apoio, há perda de um ponto para o sentar. Se, ao levantar-se, utiliza três apoios e ainda desequilibra, três pontos e

meio são subtraídos da nota máxima para o levantar. Logo, os escores finais do TSL, para tal indivíduo, seriam 4 para o sentar e 1,5 para o levantar. A nota zero é atribuída quando o avaliado só executa o ato com mais de quatro apoios ou com ajuda externa, como por exemplo, uma cadeira, uma parede ou até mesmo o próprio avaliador.

Em geral, apenas duas tentativas são necessárias para que o indivíduo consiga seu melhor resultado. Todavia, frente à possibilidade de melhoria do rendimento, o avaliador pode permitir um número maior de tentativas, muito embora isto raramente seja relevante, na prática. Na primeira execução, o avaliador deve instruir o avaliado de forma simples e direta – “Tente sentar e depois levantar de forma equilibrada, utilizando o mínimo de apoios possível”. Nas tentativas seguintes, o avaliador, baseando-se no(s) desempenho(s) anterior(es), deve fornecer informações e orientações que permitam ao avaliado melhorar seu resultado. Tal orientação é de suma importância e alguns aspectos precisam ser considerados:

- algumas pessoas têm maior facilidade na execução, se cruzarem os membros inferiores durante as ações; logo, a sugestão para fazê-lo pode melhorar o resultado;
- a chance de algum acidente fica minimizada, se for sugerido ao avaliado utilizar um ou, no máximo, dois apoios a menos que na tentativa anterior;
- a ênfase sobre a necessidade de realizar as ações com equilíbrio, evitando que o avaliado se jogue, especialmente ao sentar-se, também tende a diminuir a possibilidade de acidente.

Os melhores resultados, dentre as tentativas para o sentar e para o levantar, independentemente, são considerados para estabelecer a nota final para o teste. A nota é então representada por dois numerais (4/1,5, utilizando o exemplo dado anteriormente), sendo o primeiro relativo ao desempenho na ação de sentar e o segundo na de levantar.

Características ordinais dos escores no TSL:

Considerando as características inerentes a este tipo de escala (ordinal), o tipo de variável mensurada (destreza nas ações de sentar e levantar do solo) e a tendência à assimetria da distribuição, recursos estatísticos não-paramétricos devem ser preferencialmente utilizados na análise dos resultados do TSL, principalmente em amostras pequenas ($n < 40$) (4). Isto posto, os testes de Wilcoxon e Mann-Whitney são os indicados para compararem-se resultados entre duas amostras dependentes e independentes entre si, respectivamente. Já os testes de Friedman e Kruskal-Wallis são os equivalentes para estudos com mais de dois grupos (5,7). Relações com outras variáveis devem ser estabelecidas, a partir da correlação de postos de Spearman (4,5). Muito embora alguns autores considerem amostras com $n > 20$ suficientemente grandes e robustas para comportar o uso de recursos da estatística paramétrica, em estudos de fidedignidade (14,15), optamos por assumir uma postura mais coerente, respeitando as características da escala ordinal e não utilizamos recursos paramétricos, independentemente da magnitude da amostra. É importante res-

saltar, ainda, que diferenças de 0,5 ponto nos escores do TSL podem refletir fenômenos distintos, principalmente em estudos de fidedignidade. Exemplificando: se dois avaliadores divergem em suas notas, atribuindo 3 e 2,5 a um dado desempenho, a diferença reside na percepção ou não de desequilíbrio, pois ambos concordaram em que dois apoios extras foram utilizados. Pode-se dizer, aqui, que houve uma concordância relativa, uma vez que o maior grau de subjetividade na escala encontra-se na identificação da perda de equilíbrio. Porém, se os mesmos avaliadores atribuem 2,5 e 2,0 a uma execução, a divergência é, não só em relação à presença de desequilíbrio, como também quanto ao número de apoios utilizados (2 e 3, respectivamente). Em adendo, essa divergência é potencialmente mais séria e menos subjetiva que a apresentada, quando as notas diferem em um ponto (ex: 2,5 e 1,5 ou 5 e 4), porque, neste caso, a dúvida originou-se apenas da utilização ou não de um apoio. Sendo assim, para efeito dos estudos que seguem, definimos os seguintes termos como concordância absoluta (CA), quando ocorriam notas idênticas; como discordância relativa subjetiva (DRS), quando as notas diferiam apenas quanto à presença de desequilíbrio; como discordância relativa objetiva (DRO), quando as notas diferiam apenas pela detecção de um apoio e, como discordância importante (DI), todas as demais situações. Levando-se em consideração que a maior concordância absoluta possível seria desejável, mas que em um método parcialmente subjetivo e objetivo, o componente que envolve maior subjetividade deve resultar em diferenças mais frequentes entre avaliações ou avaliadores, arbitramos que uma proporção teórica satisfatória, consoante com a simplicidade do procedimento, para os resultados do TSL, tanto no sentar como no levantar, como sendo igual a 80% (CA), 10% (DRS), 7% (DRO) e 3% (DI) na análise dos dados.

Estudo 1 – Estudo preliminar da fidedignidade interavaliadores

A amostra consistiu de 29 adultos (10 mulheres e 19 homens), com idade variando entre 15 e 58 anos (33 ± 13), apresentando estatura e peso corporal de 168 ± 7 cm e 70 ± 17 kg, respectivamente, e Índice de Massa Corporal (IMC) igual a $24,5 \pm 6$ kg/m² (média \pm dp).

Para medir-se o peso corporal, utilizou-se uma balança Yara, com leitura de meio quilograma. A medida de estatura foi realizada através da fixação a uma parede de uma fita métrica, com leitura de meio centímetro.

Cada indivíduo foi testado em seu desempenho no sentar e levantar, avaliado simultaneamente e independentemente por dois investigadores treinados no método, estando os mesmos lado a lado, de frente para o avaliado. As instruções para eventual melhoria do desempenho, entre a primeira e a segunda tentativas, ficaram sempre a cargo do mesmo avaliador. Em nenhum momento, durante os testes, os avaliadores conversavam sobre os resultados.

Na análise dos resultados, estabeleceu-se a mediana dos escores obtidos para caracterizar a amostra, quanto ao desempenho nas ações de sentar e levantar do solo, e os percentuais de CA, DRS, DRO e DI, para avaliar a coerência dos escores atribuídos pelos avaliadores. A estatística do qui-quadrado foi então utilizada, para comparar as distribuições das frequências esperadas e obtidas. Em adendo, o teste de Wilcoxon foi utilizado para detectar diferenças significativas ($p < 0,05$) entre a tendência central das notas dos dois avaliadores. Uma elevada potência foi estimada para os testes ($P > 0,95$), baseando-se no número de indivíduos estudados e considerando a diferença mínima detectável igual a um ponto na escala, o que corresponde à divergência quanto à utilização de, pelo menos, um apoio.

Estudo 2 - Estudo das fidedignidades interavaliadores e intravaliador

A amostra foi composta por 10 indivíduos, de ambos os sexos, sendo duas crianças e oito adultos, com idade de 24 ± 10 anos (média \pm dp). Os indivíduos, que avaliados em condições normais apresentavam notas cinco e cinco no TSL, simularam todos os possíveis escores, obedecendo a uma ordem randômica estratificada, para a ação de sentar e para a de levantar, perfazendo um total de 100 movimentos para cada ação. Tendo em vista que o TSL envolve ações relativamente simples, do ponto de vista motor para um indivíduo jovem e saudável, os maiores escores tendem a ser mais comuns, principalmente em crianças e adultos. Assim, visando a simular uma situação de aplicação prática real, os desempenhos para os escores cinco e quatro e meio foram simulados 14 vezes, e o desempenho correspondente ao escore quatro, 12 vezes. Partindo desse último, em ordem decrescente, o desempenho equivalente a cada escore era simulado uma vez menos. A graduação zero foi simulada, portanto, apenas quatro vezes (Tabela I).

TABELA I. Número de simulações por escore na escala de graduação do teste (estudo 2)

	Escore										
	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
Número de simulações	4	5	6	7	8	9	10	11	12	14	14

Todos os movimentos foram filmados por uma câmera de vídeo VHS montada em um tripé, a partir de um ângulo de 45° com a visão frontal do avaliado e de uma distância que permitia observar todo o corpo do avaliado e também o solo. Alguns cuidados adicionais foram tomados, tais como: iluminação (utilizou-se a do local e também a da câmera), função *pause*, sendo utilizada ao término de cada execução até o início da seguinte. Cada ato de

sentar era sucedido pelo levantar. Antes da filmagem de cada movimento, os indivíduos eram orientados sobre como proceder para que a ordem preestabelecida dos escores fosse respeitada. Um ensaio não gravado, caso o avaliado sentisse necessidade, também era permitido. Dessa forma, em raras situações os indivíduos não obtiveram sucesso ao simular as execuções.

O filme foi, então, assistido, sem o recurso de áudio, por três avaliadores, independentemente, não sendo permitido voltar a fita após ou durante a execução de qualquer dos movimentos. Parar a fita só era possível após a conclusão de cada ato, para se decidir sobre o escore correspondente ao desempenho observado, tentando simular a situação real de aplicação do teste. As notas de cada avaliador para cada ação foram registradas pelo próprio em um formulário específico, que continha 100 seqüências de dois espaços (um para cada ação), localizados dois a dois, uma abaixo da outra.

Na análise estatística dos dados, considerou-se, para cada execução simulada, a nota mais freqüente ou a mediana, caso todas as notas divergissem, como a que melhor avaliava o desempenho. A partir disso, se não houvesse concordância absoluta entre as notas, considerou-se a maior diferença para o estabelecimento dos percentuais de CA, DRS, DRO e DI. A exemplo do estudo anterior, a estatística do qui-quadrado foi utilizada, para comparar as distribuições entre as freqüências esperadas e obtidas. O teste de Friedman foi também utilizado para verificar diferenças significativas ($p < 0,05$) entre os escores dos avaliadores.

Após 10 meses, um dos avaliadores assistiu novamente à fita, buscando verificar também a fidedignidade intravaliador, a fim de fornecer maiores subsídios sobre a consistência interna do procedimento. Os procedimentos estatísticos foram aqueles utilizados no estudo de fidedignidade interavaliadores, sendo a presença de vieses sistemáticos testadas pelo teste de Wilcoxon. Verificou-se também elevada potência para os testes ($P > 0,95$), utilizando o mesmo critério descrito no estudo preliminar de fidedignidade interavaliadores.

Estudo 3 - Estudo das fidedignidades interdias e intradia

A amostra foi composta por 10 indivíduos (4 homens e 6 mulheres), com idade entre 20 e 42 anos [(30 ± 9), média ± dp], estatura de 165 ± 7 cm e peso corporal de 74

± 17 kg (IMC = 27,2 ± 6 kg/m²), medidos em uma balança e estadiômetro Filizola, com leitura de 100 g para o peso e de 0,1 cm para a estatura.

Para a observação da variabilidade interdias, testes foram realizados, em quatro dias distintos, pelo mesmo avaliador, no período máximo de nove dias. Em um dos dias, aleatoriamente determinado para cada indivíduo, também foi testada a variabilidade em 10 medidas consecutivas do teste, com intervalo de 30 segundos entre as mesmas (Figura 2). Todas as avaliações foram feitas entre as 12 e 15 horas da tarde, sem a realização de atividade física, naquele dia.

FIGURA 2. Abordagem metodológica do estudo de estabilidade do TSL utilizando como exemplo a rotina de avaliações de um indivíduo na amostra.



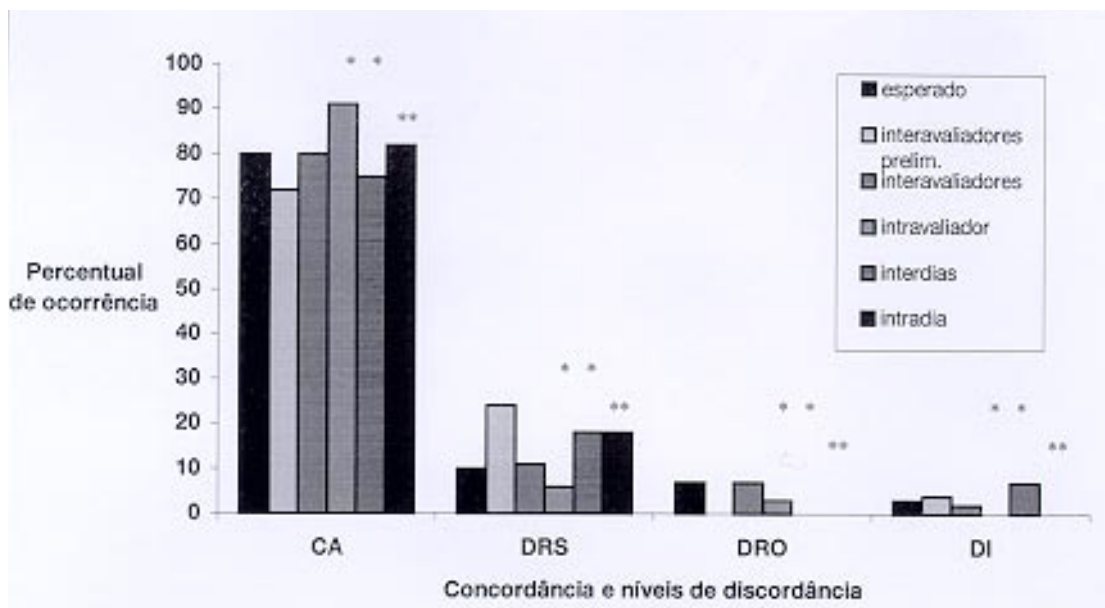
Uma vez que os escores no TSL são determinados, considerando o melhor desempenho alcançado, os dados foram analisados de forma coerente com esse pressuposto, ou seja, considerando o maior escore, dentre os quatro dias ou dentre as 10 execuções, para cada estudo, respectivamente, como o que melhor refletia o desempenho de cada avaliado. Nesse sentido, a mediana entre os melhores escores dos indivíduos foi calculada, visando a caracterizar a amostra quanto ao desempenho nas ações de sentar e levantar do solo. Isto posto, os percentuais de concordância absoluta, assim como de discordâncias relativas e de discordância importante foram estabelecidos dentro de um universo de 40 e 100 avaliações, para a fidedignidade interdias e intradia, respectivamente. Utilizou-se, posteriormente, a estatística do qui-quadrado, para comparar as freqüências obtidas com as esperadas, como nos estudos anteriores. Além disso, objetivando identificar algum efeito de fadiga ou de aprendizagem sobre o desempenho, testou-se se a freqüência com que cada escore ocorria, nas 40 e 100 avaliações, para cada estudo, diferia da freqüência com que ocorriam em cada dia ou avaliação, nos 10 indivíduos. O teste de Friedman foi utilizado para a identificação de diferenças significativas ($p < 0,05$) entre os desempenhos, em quatro dias distintos e entre as 10 medidas consecutivas. Potências seguras foram estimadas para as abordagens re-

ferentes às fidedignidades interdias ($P > 0,9$) e intradia ($P > 0,83$), baseando-se no número de indivíduos estudados e considerando qualquer modificação no número de apoios utilizados como a mínima diferença detectável.

Resultados

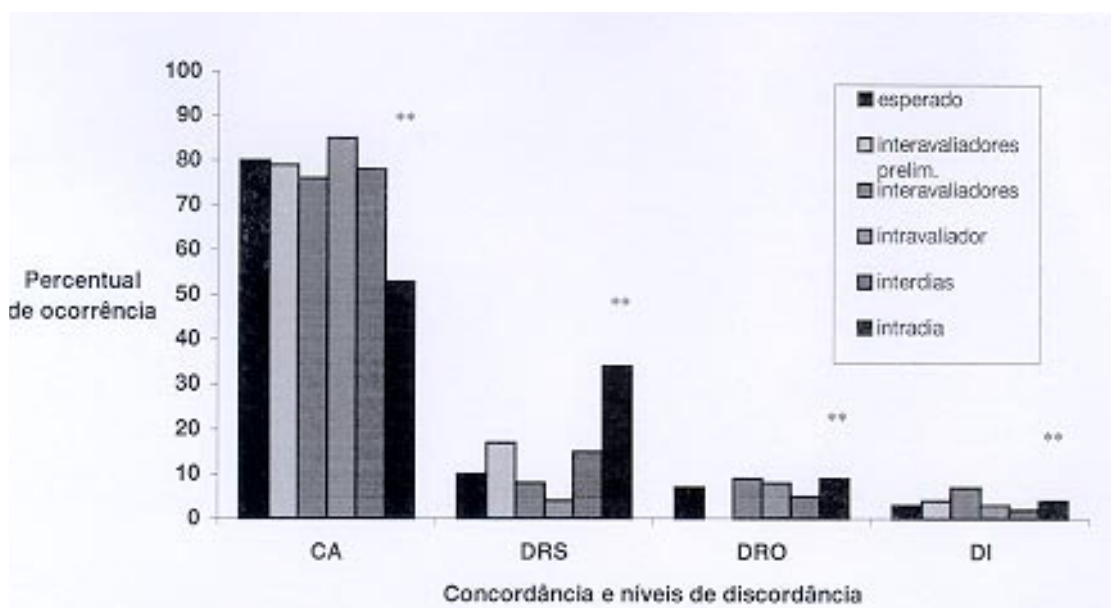
Nas figuras 3 e 4 observam-se os percentuais esperados e obtidos de CA, DRS, DRO e DI para as ações de sentar e levantar, respectivamente, em todos os estudos.

FIGURA 3. Representação da concordância e níveis de discordância esperados e obtidos em cada estudo para o sentar



CA – Concordância absoluta; DRS – Discordância relativa subjetiva; DRO – Discordância relativa objetiva; DI – Discordância importante; * diferença entre a distribuição esperada e a obtida (teste qui-quadrado, $p < 0,05$); ** diferença entre a distribuição esperada e a obtida (teste qui-quadrado, $p < 0,01$)

FIGURA 4. Representação da concordância e níveis de discordância esperados e obtidos em cada estudo para o levantar



CA – Concordância absoluta ;CRS – Discordância relativa subjetiva ;CRO – Discordância relativa objetiva; DI – Discordância importante ** diferença entre a distribuição esperada e a obtida (teste qui-quadrado, $p < 0,01$)

No estudo preliminar de fidedignidade interavaliadores (estudo 1), os escores distribuíram-se entre 3,5 e 5 (mediana = 4,5), para a ação de sentar e entre 3 e 5, para a de levantar (mediana = 4,5), o que se encontra de acordo com o usualmente observado, em nossa experiência com mais de 400 TSL realizados, para indivíduos dessa faixa etária (dados não publicados). Os avaliadores não diferiram quanto aos escores atribuídos ($p > 0,05$), verificando-se ainda CA em 72% das ações, para o sentar e em 79% das ações para o levantar. Além disso, a principal fonte de divergência era a identificação de desequilíbrio, tanto no sentar (DRS = 24%), quanto no levantar (DRS = 17%). Dessa forma, não foram evidenciadas diferenças significativas entre as proporções arbitradas como apropriadas e as obtidas de CA, DRS, DRO e DI ($p = 0,06$ para o sentar e $p = 0,34$ para o levantar).

No que concerne ao principal estudo de fidedignidade interavaliadores, 80% das execuções para o sentar foram avaliadas identicamente pelos três avaliadores ($p = 0,86$), enquanto 76% o foram para o levantar ($p = 0,82$). Novamente, a subjetividade referente ao desequilíbrio foi a maior responsável pela diferença nos escores no sentar (DRS = 11%), enquanto a diferença de até um apoio no escore de levantar (DRO = 9%). Novamente, os percentuais de distribuição apresentaram frequências próximas às esperadas para o sentar ($p = 0,93$) e para o levantar ($p = 0,09$).

Quanto à fidedignidade intravaliador, não houve diferença entre os escores nas duas avaliações para as duas ações ($p > 0,05$), tendo ainda sido verificada concordância absoluta em 91% das ações, para a ação de sentar e em 85%, para a de levantar. A concordância absoluta para o sentar foi superior à previamente arbitrada como apropriada, originando diferenças significativas entre as proporções de CA, DRS, DRO e DI, esperadas e obtidas ($p < 0,04$). Já para a ação de levantar, os resultados obtidos foram semelhantes aos esperados ($p = 0,25$). Pode-se ainda destacar o fato de que, em 9% e 15% dos casos em que não houve concordância absoluta, respectivamente para o sentar e para o levantar, observou-se apenas DRS e não DRO, ou seja, as discrepâncias ocorreram apenas na identificação subjetiva de presença ou ausência de desequilíbrio perceptível.

Na investigação sobre a estabilidade do teste (estudo 3), os escores individuais variaram entre 3,5 e 4,5, em dias distintos e entre 3,5 e 5,0, no mesmo dia, para a ação de sentar (mediana = 4,25). Na ação de levantar, os escores individuais ficaram entre 3,5 e 5,0, em dias próximos e, também, no mesmo dia (mediana = 4). Não foi evidenciada qualquer diferença significativa para as notas de sentar e levantar em dias distintos ($p = 0,5$; $p = 1,0$), como também nas 10 medidas feitas em um único dia ($p = 0,36$; $p = 0,25$, respectivamente). Verificou-se que, em quatro dias próximos, os escores para o sentar e para o levantar repetiam-se em 75% e em 77% das avaliações, respectivamente. Em várias avaliações, no mesmo dia, a estabilidade dos escores era de 82% para o sentar e de 53% para o levantar. As proporções observadas de concordância absoluta, discordâncias relativas subjetiva e objetiva e de discordância importante, diferiram das esperadas no mesmo dia, para ambas as ações ($p < 0,01$). No sentar, a ocorrência de DRO e DI (0%) foi inferior à esperada, enquanto no levantar CA ocorreu menos que o desejado (53%). Em dias distintos, houve dife-

renças entre o observado e o esperado para o sentar ($p = 0,03$), mas não para o levantar ($p = 0,7$). Além disso, a comparação entre a frequência total de cada escore na amostra, e a observada em cada dia (fidedignidade interdias) ou avaliação (fidedignidade intradia), demonstrou que as diferenças nos escores ocorriam, aleatoriamente, para a ação de sentar ($p > 0,74$ e $p > 0,34$, respectivamente) e para a de levantar ($p > 0,11$ e $p > 0,56$, respectivamente).

Discussão

Os tipos de fidedignidade necessários para o estabelecimento da consistência de um novo instrumento de medida encontram-se classicamente caracterizados na literatura (5,22). Todavia, o tratamento estatístico dos dados obtidos, ou seja, os índices de fidedignidade a serem apresentados e a interpretação dos mesmos, têm sido alvo de discussão (4,16). Exemplificando, é comum observar estudos visando a estabelecer a fidedignidade intra e inter-dias de procedimentos de medida diversos, utilizando, para tal, distintos métodos estatísticos para avaliar os dados obtidos. Como efeito, alguns ensaios acabam por tratar os dados de forma inadequada, comprometendo suas inferências. Alguns tratamentos estatísticos não são excludentes entre si, e sim fornecem informações complementares (4).

Grande parte dos estudos com as características do presente trabalho apresentam os resultados dos testes de hipótese (Wilcoxon, teste-t, Friedman e ANOVA, entre outros) informando que os escores de avaliações ou avaliadores distintos, para uma mesma amostra, tendem ou não a se manter estáveis. Esses recursos estatísticos apenas podem informar se existem erros sistemáticos entre as avaliações (efeitos de aprendizagem ou fadiga sobre o desempenho, por exemplo), não sendo capazes de fornecer informações sobre a frequência de erros randômicos, aqueles atribuíveis a variações biológicas, mecânicas ou limitações do próprio protocolo. Outro aspecto que merece destaque é que a potência desses testes é diretamente proporcional à homogeneidade e tamanho da amostra. Assim, em estudos de fidedignidade com amostras heterogêneas, com desempenhos diversos quanto à variável avaliada, há menor probabilidade dos testes de hipótese identificarem erros sistemáticos. Isso, porque grande parte da variabilidade de respostas se deve a erros randômicos. Dessa forma, os testes acima mencionados são úteis, mas não devem ser utilizados isoladamente no estabelecimento da fidedignidade, uma vez que fornecem dados apenas de uma fidedignidade relativa, isto é, dependente da homogeneidade da amostra (4).

Outros estudos estatisticamente mais sofisticados fornecem, ainda, coeficientes de fidedignidade, dentre os quais o mais comumente utilizado é o coeficiente de correlação intraclasses, derivado da ANOVA com medidas repetidas (5,22). O coeficiente *kappa* é o equivalente para variáveis ordinais (2). Entretanto, a utilização de tais coeficientes, em estudos de fidedignidade, não depende somente da escala de medida que, no caso do TSL, é ordinal, mas também do desempenho que cada escore representa. Isto posto, não se calculou o coeficiente *kappa* nos estudos de fidedignidade do TSL porque, como descrito na seção de metodologia, diferenças de 0,5 ponto, como por exemplo

entre as notas 3,5 e 3,0, compreendem divergências no número de apoios utilizados e também na presença de desequilíbrio. Comparativamente, uma maior variação no escore, equivalente a de um ponto, por exemplo, reflete menor grau de discordância no julgamento do desempenho. Assim sendo, o coeficiente trataria todas as diferenças de 0,5 ponto como representando diferenças idênticas no desempenho, tendenciando erroneamente a análise. Optou-se, então, por desenvolver uma análise com características um pouco mais qualitativas e, provavelmente, mais aplicáveis em estudos futuros. A observação da frequência com que a concordância absoluta, discordâncias relativas e discordância importante ocorriam, em relação a um pressuposto teórico arbitrário de distribuição das mesmas, possibilitou que as peculiaridades da escala do TSL fossem contempladas. Além disso, as elevadas potências observadas conferem alta probabilidade de identificação de erros sistemáticos, na análise dos grupos de dados por avaliadores ou avaliações, apontados pelos testes de hipótese (Wilcoxon e Friedman). Todos esses aspectos reunidos reforçam a propriedade do desenho metodológico, utilizado no presente estudo.

Os resultados do estudo 1 apontam que o TSL fornece escores com poucos erros sistemáticos entre avaliadores, em crianças e adultos jovens e saudáveis, com desempenhos nos graus superiores da escala. Porém, sendo uma amostra reduzida e homogênea, quanto à destreza nas ações de sentar e levantar do solo, não seria possível extrapolar tal fidedignidade para os desempenhos que perfazem os níveis inferiores e medianos da escala. Descontando-se os percentuais de ocorrência de CA e DRS, somados, observa-se que discordâncias importantes ou aquelas que envolvem a utilização de apoios ocorreram em menos de cinco, para cada 100 avaliações, realizadas nas ações de sentar e levantar.

Os resultados do estudo 2 trazem informações que refletem melhor a fidedignidade interavaliadores, uma vez que se originam de uma frequência estratificada dos escores. Normalmente, coeficientes de correlação intraclasse e coeficientes *kappa*, relacionados à fidedignidade interavaliadores, variam desde modestos 0,66 a excelentes 0,94, com raros estudos evidenciando valores bem próximos à unidade, em instrumentos desenvolvidos para a medida da funcionalidade de pessoas não saudáveis (23,16) e de equilíbrio e força de membros inferiores (10). Todavia, apesar das características dos escores no TSL não possibilitarem a identificação confiável e coerente de tais coeficientes, o alto grau de concordância, observado para o sentar e para o levantar, atesta uma elevada fidedignidade. Percentuais adicionados de CA, DRS e DRO, as duas últimas compreendendo as menores diferenças possíveis de ordem subjetiva e objetiva na escala, respectivamente, são os mais frequentes, totalizando cerca de 93% das avaliações, para o levantar e 98%, para o sentar. Em outras palavras, implica dizer que, em um universo de 100 avaliações, apenas 2 a 7, independentemente da ação em questão, podem ser fruto de um erro importante do avaliador. Já em relação à fidedignidade intravaliador, observou-se consistência extremamente alta nos dados relativos ao sentar, em comparação ao levantar, embora ambos apresentassem percentuais de ocorrência iguais ou inferiores a três, para discordâncias de mais

de um ponto na escala de medida, entre o mesmo avaliador, o que é bastante apropriado para um teste com essas características.

A exemplo do estudo 2, o estabelecimento da estabilidade, ou seja, das fidedignidades inter e intradia, de instrumentos de medida similares ao TSL também originou coeficientes de razoáveis a excelentes, como para a avaliação da força e 'funcionalidade física' em pessoas idosas, por exemplo, nos quais coeficientes de correlação intraclasse de 0,72 a 0,97, foram evidenciados (11,12,21). Aqui também a soma dos percentuais de CA e DRS, apresentou maior frequência do que a esperada. Parece, realmente, que o desempenho tende a não variar muito em avaliações próximas e, ainda, que a variação tende a ser aleatória, minimizando a possibilidade de aprendizado ou fadiga pela realização do teste. As variações evidenciadas nos escores parecem ser fruto de variações individuais no desempenho.

Não obstante, parece-nos ainda necessário investigar a estabilidade do teste em populações com características específicas distintas, tais como crianças, idosos e mulheres. O nível de treinamento físico talvez exerça influência e, assim, pode haver diferença nas fidedignidades inter e intradia do teste para homens e mulheres treinados, em comparação a sedentários. Em adendo, as características do treinamento, ou seja, se o mesmo é direcionado para a melhoria primordialmente da força ou flexibilidade ou, ainda, da resistência muscular, bem como os grupamentos musculares enfatizados (membros inferiores, superiores ou ambos) também podem originar tendências distintas nos parâmetros de fidedignidade.

Contudo, de forma geral, pode-se dizer que a fidedignidade interdias tende a ser superior à intradia em adultos saudáveis, como pode ser evidenciado, pelos percentuais de concordância e dos níveis de discordância, principalmente em relação à ação de levantar. Outro ponto de interesse é que o tempo mínimo de intervalo entre avaliações em um mesmo dia (30 segundos), adotado neste estudo, parece ser apropriado para tentativas sucessivas do TSL, já que não foi observado erro sistemático nas medidas, como seria esperado, caso houvesse algum efeito de fadiga ou de aprendizagem significativo. Assim, a menor consistência nos resultados em avaliações, realizadas no mesmo dia, em comparação às realizadas isoladamente, em dias próximos, deve ser fruto da perda de motivação, a partir da realização seguida dos movimentos.

Em todos os estudos houve uma tendência para uma maior consistência para o ato de sentar do que para o de levantar. Pelos resultados dos testes de hipótese, não houve presença significativa de erros sistemáticos em todos os desenhos metodológicos abordados, cuja razão parece residir na presença de erros randômicos. Entende-se aqui, principalmente, a influência de características biológicas. Mesmo assim, é prudente observar que tal influência tende a ser bem reduzida, senão desprezível, de acordo com os percentuais de ocorrência de concordância nos escores e os resultados dos testes de hipótese. Nesse sentido, ainda é relevante comentar que a principal fonte de discordância diz respeito à percepção de desequilíbrio, ou seja, 0,5 ponto. Logo, erros na medida de origem sistemática e/ou randômica tendem a resultar em variação mínima, dentro

da escala de mensuração. Além disso, a necessidade de utilização de mais um apoio nas ações do TSL pode significar maiores limitações físicas do avaliado que a presença de um desequilíbrio perceptível. A simplicidade na realização e na avaliação dos resultados no TSL provavelmente justifica os baixos percentuais de discordâncias importantes observados.

Conclusão

A partir dos quatro estudos desenvolvidos foi possível identificar importantes características da fidedignidade do TSL. Sendo assim, algumas conclusões e implicações podem ser estabelecidas.

O TSL mede a destreza nas ações de sentar e levantar do solo. Tal destreza tende a ser inversamente proporcional à necessidade de apoio no solo e à perda de equilíbrio, quando da execução das ações. Considerando que o TSL pode medir diretamente esse desempenho, não parece haver dúvidas quanto à sua validade lógica (5,22). O fato de que tais atos envolvem, até certo ponto, força e potência musculares, bem como flexibilidade de membros inferiores, confere ao teste características a serem exploradas em avaliações funcionais e na prescrição de exercícios. Todavia, estudos futuros para a identificação das variáveis potencialmente intervenientes, nas ações de sentar e levantar, precisam ser conduzidos, a fim de fornecerem as informações necessárias sobre quais inferências podem ser traçadas, a partir dos resultados do teste. A sensibilidade do teste a modificações provindas do treinamento com exercícios físicos e a estabilidade do mesmo, em outras populações, também devem ser pesquisadas.

O TSL é um instrumento que apresenta elevada simplicidade em sua aplicação e graduação, sendo fidedigno, independentemente da magnitude do escore ou desempenho apresentado. Nas poucas oportunidades em que os avaliadores diferem em seu julgamento, a diferença tende a ser mínima ao longo da escala do teste, ou seja, de 0,5 ponto, o que corresponde, principalmente, à percepção de desequilíbrio.

O TSL também se mostrou estável em uma amostra de indivíduos saudáveis, relativamente heterogênea quanto à idade, sexo e IMC. Talvez em amostras homogêneas, sua estabilidade seja ainda maior. Sendo assim, o procedimento deve ser sensível a adaptações provindas do treinamento, principalmente para aqueles indivíduos que apresentarem desempenhos limitados, nas ações de sentar e levantar do solo.

É possível concluir, ainda, que a utilização do TSL em escolas, centros de atividades, como clubes, academias e consultórios médicos pode consistir em uma estratégia útil, para identificar os indivíduos que possuem maiores limitações no desempenho de atividades simples como o sentar e o levantar.

Referências Bibliográficas

1. ALEXANDER, N.B. et al. Rising from the floor in older adults. *J Am Geriatr Soc.* 1997;45:564-569.
2. ALTMAN, D.G. *Practical statistics for medical research.* London, Chapman & Hall, 1997.
3. ARAÚJO, C.G.S. Teste de sentar-levantar - apresentação preliminar de um procedimento para avaliação em Medicina do Exercício e do Esporte. *Revista Brasileira de Medicina do Esporte.* 1999;5:179-182.
4. ATKINSON, G., NEVILL, A.M. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med.* 1998;26:217-238.
5. BERG, K.E., LATIN, R.W. *Essentials of modern research methods in health, physical education, and recreation.* Englewood Cliffs – NJ, Prentice-Hall, 1994.
6. BLAIR, S.N. et al. Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women. *JAMA.* 1996;276:205-210.
7. CHAN, Y., WALMSLEY, R.P. Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Phys Ther.* 1997;77:1755-1761.
8. CHENG, P.T. et al. The sit-to-stand movement in stroke patients and its correlation with falling. *Arch Phys Med Rehabil.* 1998;79:1043-1046.
9. EVANS, W.J, ROUBENOFF, R., SHEVITZ, A. Exercise and the treatment of wasting: aging and human immunodeficiency virus infection. *Semin Oncol.* 1998;25 (Suppl. 6):112-122.
10. FRANCHIGNONI, F. et al. Reliability of four simple, quantitative tests of balance and mobility in healthy elderly females. *Aging (Milano).* 1998;10:26-31.
11. GERETY, M.B. et al. Development and validation of a physical performance instrument for the functionally impaired elderly: the Physical Disability Index (PDI). *J Gerontol.* 1993;48:M33-M38.
12. JONES, C.J., RIKLI, R.E., BEAM, W.C. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Res Q Exerc Sport.* 1999;70:113-119.
13. MAZZEO, R.S. et al. Exercise and physical activity for older adults. *Med Sci Sports Exerc.* 1998;30:992-1008.
14. MOTULSKY, H.J. *Intuitive Biostatistics.* London, Oxford University Press, 1995.

15. OTTENBACHER, K.J., TOMCHECK, S.D. Measurement variation in method comparison studies: an empirical examination. *Arch Phys Med Rehabil.* 1994;75:505-512.
16. OTTENBACHER, K.J. et al. The reliability of the Functional Independence Measure: a quantitative review. *Arch Phys Med Rehabil.* 1996;77:1226-1232.
17. PAFFENBARGER Jr., R.S. et al. The association of changes in physical-activity level and other lifestyle characteristics with mortality among men. *N Engl J Med.* 1993;328:538-545.
18. POLLOCK, M.L., EVANS, W.J. Resistance training for health and disease: introduction. *Med Sci Sports Exerc.* 1999;31:10-11.
19. POLLOCK, M.L. et al. The recommended quantity and quality of exercise for developing and maintaining cardiorespiratory and muscular fitness, and flexibility in healthy adults. *Med Sci Sports Exerc.* 1998;30:975-991.
20. RILEY, P.O., KREBS, D.E., POPAT, R.A. Biomechanical analysis of failed sit-to-stand. *IEEE Trans Rehabil Eng.* 1997;5:353-359.
21. ROORDA, L.D. et al. Measuring functional limitations in rising and sitting down: development of a questionnaire. *Arch Phys Med Rehabil.* 1996;77:663-669.
22. THOMAS, J.R., NELSON, J.K. Research methods in physical activity. Champaign – IL, Human Kinetics, 1996.
23. WINOGRAD, C.H. et al. Development of a physical performance and mobility examination. *J Am Geriatr Soc.* 1994;42:743-749.

Agradecimentos

Os autores agradecem aos Professores Carla Werlang Coelho e José Ricardo Vianna pelo auxílio na coleta de dados e na graduação de desempenhos, em alguns dos estudos aqui reportados. Os autores também agradecem à Professora Denise Sardinha Mendes Soares de Araújo, pela ajuda na coleta dos dados e pelos comentários críticos sobre este texto e ao Professor Wallace David Monteiro, pelas sugestões e críticas.