J Rehabil Med 2017; 49: 677–681

ORIGINAL REPORT

Check for updates

REPRODUCIBILITY OF CLINICIAN-FRIENDLY PHYSICAL PERFORMANCE MEASURES IN INDIVIDUALS WITH OBESITY

Nicola A. MAFFIULETTI, PhD[1], Gabriella TRINGALI, MSc[2], Alessandra PATRIZI, MSc[2], Fiorenza AGOSTI, MSc[2] and Alessandro SARTORIO, MD[2,3]

From the [1]Human Performance Laboratory, Schulthess Clinic, Zurich, Switzerland, [2]Experimental Laboratory for Auxo-endocrinological Research, Istituto Auxologico Italiano, Verbania and Milan, [3]Division of Auxology and Metabolic Diseases, Istituto Auxologico Italiano, Verbania, Italy

Objective: **To evaluate the reproducibility (reliability and agreement) of different physical performance measures in individuals with obesity.**

Methods: **Forty subjects (20 men, 20 women), mean age 29 years, mean body mass index (BMI) 42 kg/m² completed several clinician-friendly performance-based tests (walking, stair-climbing, sit-to-stand, static balance, flexibility and strength) on 2 different occasions (test-retest design). Intraclass correlation coefficients (reliability) and smallest detectable changes (agreement) were calculated for each outcome measure.**

Results: **Intraclass correlation coefficients were relatively high (range 0.84–0.94) for all the performance-based measures (i.e. acceptable reliability). Smallest detectable changes were overall quite high and beyond the arbitrarily-defined minimal clinically important changes (i.e. poor agreement) for 3 out of 8 variables (sit-to-stand time, time-in-balance with eyes closed, and sit-and-reach distance).**

Conclusion: **The clinician-friendly performance-based tests for individuals with obesity considered in this study appear legitimate for discriminative purposes, such as in cross-sectional studies. However, for longitudinal assessments (evaluative purposes), some measures should be used with greater caution due to limited agreement. Careful consideration should be given to the evaluation of physical performance in people with obesity, particularly in the context of conservative or surgical treatment for weight loss.**

Key words: obesity; physical performance; outcomes; reproducibility.

Accepted Jun 15, 2017; Epub ahead of print Aug 9, 2017

J Rehabil Med 2017; 49: 677–681

Correspondence address: Nicola A. Maffiuletti, Schulthess Clinic, Lengghalde 2, CH-8008 Zurich, Switzerland. E-mail: nicola.maffiuletti@kws.ch

Obesity results in functional limitations in terms of restrictions in performing fundamental physical actions used in daily life (1). In addition to self-report measures of physical functioning, performance-based assessments are often conducted longitudinally in individuals with obesity to substantiate the effects of conservative or surgical treatment for weight loss (2–6). The most common functional evaluations in the clinical setting entail activities of daily living, such as walking, stair-climbing, postural transitions and static balance, but also components of physical fitness, such as flexibility and muscle strength. Thus, walking speed (3), timed tests (stair, sit-to-stand, balance) (4, 6), sit-and-reach distance (2), and 1-repetition maximum load (5) are frequently assessed in individuals with obesity. Compared with laboratory-based assessments with costly and complex equipment, these evaluations are extremely accessible, inexpensive and easy to perform for both the clinician and the patient. However, interpretation of such evaluations is often complicated by the absence of reproducibility and normative data in this specific population, which precludes the opportunity of determining whether a difference/change in performance is meaningful.

To the best of our knowledge, the measurement properties (such as reliability and agreement) of the multitude of performance-based tests for adults with obesity have not yet been studied adequately, which seriously affects the validity of these evaluations. In particular, the smallest detectable change (SDC), also referred to as minimal detectable change, i.e. the smallest amount of change that can be detected by a measure that corresponds to a noticeable change in ability (7), is particularly important, as it allows the clinician to differentiate between true changes and changes due to measurement error.

The aim of this study was to determine the between-day reproducibility of several clinician-friendly physical performance measures in individuals with obesity. More specifically, a test-retest design was implemented to examine the reliability and agreement (SDC) of selected walking, stair-climbing, sit-to-stand, balance, flexibility and strength tests that are frequently administered to people with obesity. The findings of this methodological study should encourage clinicians and researchers to more consistently select tests to measure performance-based differences/changes in this population.

METHODS

Subjects

Based on the recommendations of Walter et al. (8), 39 participants would be required to statistically conclude (with 80% power) that an intraclass correlation coefficient (ICC) for test-

doi: 10.2340/16501977-2263

retest reliability is at least 0.6 (minimal) for a true ICC of 0.8 (desired). Thus, 40 subjects with obesity were included in the study (20 men and 20 women; mean age and standard deviation (SD): 29 years (SD 9); height: 170 cm (SD 10); BMI: 42 kg/m^2 (SD 4)). The main inclusion criteria were age ≥ 18 years, BMI > 30 kg/m^2, absence of severe and uncontrolled hypertension, overt uncompensated diabetes, and any major disease. The study was approved by the ethics committee of the Italian Institute of Auxology (number 01C007; acronym RIPFUN), and all participants provided written informed consent.

Experimental protocol

All subjects were admitted to the Division of Metabolic Diseases III (Italian Institute of Auxology, Piancavallo, Italy) for a multidisciplinary weight reduction programme that started 2–3 days after admission. Subjects who volunteered to participate in this study were invited to attend 2 testing sessions separated by 2–3 days (no treatment was offered during this time period). Each testing session lasted approximately 1 h and was scheduled at the same time of the day on an individual basis. Following a standardized 10-min warm-up (stationary cycling at 40–80 W), 8 different physical performance tests were randomly presented, with 2 series per test (also randomly presented). The first series was consistently used for familiarization purposes, while the actual assessments were conducted on the second series. Test series were separated by 3–5-min rest periods. Subjects wore their own shoes and received consistent verbal encouragement during testing. They were also asked to avoid exhausting exercise the day before the assessments. All tests were conducted in a laboratory and in the adjacent public space (e.g. corridors and stairs). The examiner had previous experience with the different physical performance tests used in this study.

Assessments

Walking tests. Subjects completed 2 walking tests (normal and fast pace) for ambulatory ability, according to the protocol described by Browning et al. (9). They were asked to walk back and forth 6 times along a 70-m walkway at both normal and fast speeds. They were carefully instructed to walk at their "comfortable" and "faster than normal" walking paces. For each condition, a series of 6 trials was completed. A stopwatch was used to record the time over the middle 50 m of each trial and normal and fast walking speeds were calculated as the mean of the last 5 trials per condition.

Timed stair test. Subjects completed the timed stair test for stair-climbing ability, according to the protocol described by Perron et al. (10). They were asked to stand up from a chair and walk 3 m, ascend a staircase, turn around and descend stairs, walk back to the chair, turn and sit down. The staircase comprised 13 stairs (height 15 cm, depth 32 cm); the chair had a backrest and armrests and was of standard dimensions (height 42 cm, depth 34 cm). Subjects were instructed to look forward, rest their trunk on the backrest, wait for the starting signal, walk at a comfortable pace and assume the same sitting position when returning to the chair. A stopwatch was used to record the total duration of the timed stair test.

Sit-to-stand test. Subjects completed the sit-to-stand test for transition ability, according to the protocol described by Guralnik et al. (11). They were asked to perform 5 consecutive chair rises at a comfortable pace. Their hands were folded in front of the chest with feet flat on the floor. The chair was of standard dimensions (height 42 cm, depth 34 cm). A stopwatch was used to record the total duration of the sit-to-stand test. Timing began with the command "go" and ended when the buttocks contacted the chair after the fifth trial.

Static balance tests. Subjects completed 2 static single-leg balance tests (eyes open and closed) for postural stability, according to the protocol described by Frändin et al. (12). They were asked to stand on 1 leg at a time, with the hands on their hips and with their eyes open or closed. The position of the non-weight-bearing leg was self-selected by each subject. The time until balance was lost (maximum 30 s) was recorded with a stopwatch. Three trials per leg and condition were allowed, and the mean time-in-balance with both eyes open and closed was calculated.

Sit-and-reach test. Subjects completed the sit-and-reach test for flexibility, according to the protocol described by Chen et al. (13). A flexible measuring tape was fixed to an exercise mat and subjects sat on the mat with their knees extended and the tape between their legs. Subjects were asked to reach forward slowly, as far as possible, with their hands overlapped and to hold the end position for 2 s. Three trials were allowed and the mean farthest point reached with the fingertips, using the level of the heels as recording zero (so that any measure that did not reach the heels was negative and any measure beyond the heels was positive), was the attained sit-and-reach distance.

Strength test. Subjects completed the 1-repetition maximum test for lower-extremity strength, according to the protocol described by Maffiuletti et al. (5). They were seated on a conventional horizontal leg-press machine (Technogym, Gambettola, Italy) and were asked to slowly extend their lower extremities against a pre-defined load. Subjects completed 3–4 series of 15–20 repetitions with submaximal loads, which were increased progressively, as defined by the researcher. The 1-repetition maximum load (i.e. the theoretical load that can be lifted only once) was estimated with the formula proposed by Brzycki (14).

Statistical analyses

The normality of the data was checked with Shapiro–Wilk tests. Changes in the mean between the 2 test sessions were analysed with 2-tailed paired *t*-tests to assess the presence of systematic bias. The between-day reproducibility of the different performance-based outcomes was assessed as reliability and agreement. Reliability was evaluated using intraclass correlation coefficients (ICC) with a 2-way random effects model (2, 1). An ICC > 0.70 with the lower limit of the confidence interval > 0.60 was considered acceptable (15). Agreement was evaluated using the SDC at 90% confidence interval (16) (SDC = 1.645 × √2 × standard error of measurement (SEM), where SEM = standard deviation (SD) of the difference between test sessions/√2) (17). SDC was also expressed as a percentage of mean values to produce unitless indicators and allow for comparisons. The thresholds for acceptable SDC (i.e. SDC smaller than the respective minimal important change) were arbitrarily defined for each outcome, based on respective percentage changes reported in people with obesity following an intervention (18): 11.6% for normal and fast walking speeds (3), 14.3% for timed stair test (4), 8.0% for sit-to-stand time (4), 20.5% for time-in-balance (eyes open and closed) (6), 47.0% for sit-and-reach distance (2), 37.4% for 1-repetition maximum load (5). For all the analyses, a *p*-value below 0.05 was considered statistically significant. All statistical analyses were performed with the Statistica software (Statistica 7.0, Statsoft, Tulsa, OK, USA).

Table I. Physical performance measures by test session

	Test Mean (SD)	Re-test Mean (SD)	Change in the mean
Normal walking speed, km/h	5.1 (0.6)	5.2 (0.7)	0.07
Fast walking speed, km/h	6.2 (0.6)	6.2 (0.8)	0.03
Timed stair test, s	17.7 (2.6)	17.2 (3.1)	−0.50*
Sit-to-stand time, s	13.1 (3.3)	11.7 (3.2)	−1.41*
Time-in-balance eyes open, s	26.2 (7.6)	26.0 (7.7)	−0.22
Time-in-balance eyes closed, s	8.4 (6.5)	9.2 (7.3)	0.81
Sit-and-reach distance, cm	−13.5 (6.7)	−11.4 (7.0)	2.13*
1-repetition maximum load, kg	153.4 (71.5)	174.4 (76.9)	21.0*

*Significant difference (systematic bias) between test sessions ($p < 0.05$).

RESULTS

Table I shows the mean data by test session as well as respective changes in the mean. A systematic bias was observed for the timed stair test, sit-to-stand time, sit-and-reach distance and 1-repetition maximum load, with mean improvements of 2.9%, 11.3%, 17.5% and 12.8%, respectively, from test to retest.

Table II shows ICC with 95% confidence intervals (reliability) and SDC in absolute and percentage units (agreement). Reliability was acceptable for the ensemble of the performance-based measures, and the highest ICC were observed for time-in-balance (both eyes open and closed), normal walking speed and sit-to-stand time. Agreement was acceptable for walking speeds (both normal and fast), timed stair test and time-in-balance eyes open, while percentage SDCs were higher than the arbitrarily-defined minimal important changes (i.e. poor agreement) for sit-to-stand time, time-in-balance eyes closed and sit-and-reach distance.

DISCUSSION

The present methodological study establishes the reproducibility of several physical performance measures in adult individuals with obesity. The main findings were that, despite between-day reliability being found acceptable for the different performance-based outcomes (as witnessed by the relatively high ICC), systematic bias and poor agreement were observed for some of the measures (as witnessed by the relatively high SDC), thus indicating that not all the performance-based assessments can be used with confidence in subjects with obesity.

Table II. Reproducibility of physical performance measures

	$ICC_{2,1}$ (95% CI)	SDC (%)
Normal walking speed, km/h	0.906 (0.829–0.949)	0.45 (8.6)
Fast walking speed, km/h	0.859 (0.749–0.923)	0.57 (9.1)
Timed stair test, s	0.880 (0.784–0.935)	2.10 (12.0)
Sit-to-stand time, s	0.916 (0.847–0.955)	2.22 (17.9*)
Time-in-balance eyes open, s	0.943 (0.895–0.969)	4.20 (16.1)
Time-in-balance eyes closed, s	0.909 (0.835–0.951)	4.56 (52.0*)
Sit-and-reach distance, cm	0.836 (0.711–0.910)	6.30 (50.7*)
1-repetition maximum load, kg	0.862 (0.654–0.937)	61.8 (37.7)

*Percentage SDC exceeding the minimal important change (poor agreement). ICC: intraclass correlation coefficient; 95% CI: 95% confidence interval; SDC: small detectable change.

Reproducibility concerns the degree to which repeated measurements (test-retest) provide similar results in stable persons (19). The current study made a distinction between reliability and agreement, as recommended by Terwee et al. (19). Reliability concerns the degree to which subjects can be distinguished from each other, despite measurement error, and is particularly important for discriminative purposes, such as in cross-sectional studies (e.g. for comparing individuals with different degrees of obesity). The relatively high ICCs observed in our study (range 0.84–0.94) confirm that the use of the performance-based assessments appear legitimate for these purposes, although no attempt was made to examine their discriminant validity (due to the relatively small sample size). Our ICC are comparable to those previously reported across a variety of populations, including patients and older individuals (15, 20–22), thus obesity does not seem to influence the test-retest reliability of the different physical performance measures. On the other hand, agreement concerns the absolute measurement error (how close the results are on repeated measurements) and is particularly relevant for evaluative purposes in which one wants to distinguish clinically important changes from measurement error. In the present study, measurement error was estimated from SDC (i.e. the smallest change that can be interpreted as a "real" change above measurement error) (17), while the minimal clinically important change was arbitrarily defined for each outcome based on respective percentage changes reported in people with obesity following an intervention (18). Thus, SDC were overall quite close to the clinically important changes, and for 3 out of 8 outcome measures the former were higher than the latter (i.e. poor agreement). These results suggest that, contrary to the results obtained in different populations of patients without obesity and elderly individuals (7, 23, 24), but in agreement with the study of Goldberg et al. (25) on single-leg stance time, some performance-based tests should not be used (sit-to-stand, balance eyes closed and sit-and-reach) while the other assessments should be used with caution for evaluative purposes in individuals with obesity.

The presence of a systematic bias for half of the outcome measures (timed stair test, sit-to-stand time, sit-and-reach distance and 1-repetition maximum load), despite being somewhat associated with agreement results, confirmed the occurrence of a significant improvement from test to retest, which was probably due to a learning effect. These changes were quite substantial (e.g. +17% for the sit-and-reach distance), and even if they can be eventually reduced by additio-

nal familiarization and practice (an extra orientation session is probably required in longitudinal studies), learning effects should seriously be considered in the functional evaluation of individuals with obesity. Additional sources of error may include inconsistencies caused by the physical or mental status of the tested subject (e.g. a person with obesity can develop positive or negative self-esteem throughout repeated evaluations), variations in the testing procedure, or tester error. Maintaining consistency and using standardized evaluation protocols, such as using the same tester, set-up, testing order, and time of day, can result in more reliable assessments (22).

This methodological study considered clinician-friendly physical performance tests requiring non-technical, readily available, inexpensive and non-computerized equipment (a chair, a stopwatch, a measuring tape and a conventional leg press machine). In addition, all the evaluations consisted of daily-living, submaximal and comfortable efforts (including the lower extremity strength test) with no body-mounted instruments, which are certainly better tolerated by individuals with obesity than maximal-intensity and complex laboratory assessments. A preference for such simple measures, which can be performed in most settings, was made because they are more likely to be implemented in both research and clinical practice.

This study has some limitations. More psychometric qualities should be evaluated before drawing conclusions on the appropriateness of these tests in individuals with obesity. For example, no attempt was made to examine the discriminant validity (e.g. differences between individuals with moderate and severe obesity), convergent validity (e.g. relation with self-report measures of physical function) and inter-rater reliability (e.g. expert vs non-expert examiner) of performance-based outcome measures in subjects of different age and sex. Also, the thresholds for acceptable SDC were obtained from different studies and their definition can appear arbitrary; however, it is a valid option for examining the minimal clinically important change (18).

In conclusion, greater caution should be exercised when interpreting physical performance outcomes obtained in individuals with obesity. While the different walking, stair-climbing, sit-to-stand, static balance, flexibility and strength tests considered here can be used with confidence for discriminative purposes (based on acceptable reliability estimates), due to poor agreement careful consideration must be given to the use of other outcomes (time-in-balance with eyes closed, and sit-and-reach distance in particular) for evaluative purposes in longitudinal studies. The recommendations made in this paper may be helpful in monitoring differences/changes in physical performance and assessing the effectiveness of different interventions in individuals with obesity.

The authors declare no conflicts of interest.

REFERENCES

1. Verbrugge LM, Jette AM. The disablement process. Soc Sci Med 1994; 38: 1–14.
2. Benetti FA, Bacha IL, Garrido Junior AB, Greve JM. Analyses of balance and flexibility of obese patients undergoing bariatric surgery. Clinics (Sao Paulo) 2016; 71: 78–81.
3. Hortobagyi T, Herring C, Pories WJ, Rider P, Devita P. Massive weight loss-induced mechanical plasticity in obese gait. J Appl Physiol (1985) 2011; 111: 1391–1399.
4. Lyytinen T, Liikavainio T, Paakkonen M, Gylling H, Arokoski JP. Physical function and properties of quadriceps femoris muscle after bariatric surgery and subsequent weight loss. J Musculoskelet Neuronal Interact 2013; 13: 329–338.
5. Maffiuletti NA, Agosti F, Marinone PG, Silvestri G, Lafortuna CL, Sartorio A. Changes in body composition, physical performance and cardiovascular risk factors after a 3-week integrated body weight reduction program and after 1-y follow-up in severely obese men and women. Eur J Clin Nutr 2005; 59: 685–694.
6. Sartorio A, Lafortuna CL, Conte G, Faglia G, Narici MV. Changes in motor control and muscle performance after a short-term body mass reduction program in obese subjects. J Endocrinol Invest 2001; 24: 393–398.
7. Ries JD, Echternach JL, Nof L, Gagnon Blodgett M. Test-retest reliability and minimal detectable change scores for the timed "up & go" test, the six-minute walk test, and gait speed in people with Alzheimer disease. Phys Ther 2009; 89: 569–579.
8. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998; 17: 101–110.
9. Browning RC, Baker EA, Herron JA, Kram R. Effects of obesity and sex on the energetic cost and preferred speed of walking. J Appl Physiol (1985) 2006; 100: 390–398.
10. Perron M, Malouin F, Moffet H. Assessing advanced locomotor recovery after total hip arthroplasty with the timed stair test. Clin Rehabil 2003; 17: 780–786.
11. Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol 1994; 49: M85–94.
12. Frandin K, Sonn U, Svantesson U, Grimby G. Functional balance tests in 76-year-olds in relation to performance, activities of daily living and platform tests. Scand J Rehabil Med 1995; 27: 231–241.
13. Chen CN, Chuang LM, Wu YT. Clinical measures of physical fitness predict insulin resistance in people at risk for diabetes. Phys Ther 2008; 88: 1355–1364.
14. Brzycki M. Strength testing: predicting a one-rep max from reps-to-fatigue. J Health Phys Educ Rec Dance 1993; 64: 88–90.
15. Terwee CB, Mokkink LB, Steultjens MP, Dekker J. Performance-based methods for measuring the physical function of patients with osteoarthritis of the hip or knee: a systematic review of measurement properties. Rheumatology (Oxford) 2006; 45: 890–902.
16. Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. Phys Ther 2006; 86: 735–743.
17. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006; 59: 1033–1039.
18. Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, et al. Inter-rater agreement and reliability of

the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC Med Res Methodol 2010; 10: 82.
19. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007; 60: 34–42.
20. Faria CD, Teixeira-Salmela LF, Neto MG, Rodrigues-de-Paula F. Performance-based tests in subjects with stroke: outcome scores, reliability and measurement errors. Clin Rehabil 2012; 26: 460–469.
21. Hars M, Herrmann FR, Trombetti A. Reliability and minimal detectable change of gait variables in community-dwelling and hospitalized older fallers. Gait Posture 2013; 38: 1010–1014.
22. Steffen T, Seney M. Test-retest reliability and minimal detectable change on balance and ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease rating scale in people with parkinsonism. Phys Ther 2008; 88: 733–746.
23. Boer PH, Moss SJ. Test-retest reliability and minimal detectable change scores of twelve functional fitness tests in adults with Down syndrome. Res Dev Disabil 2016; 48: 176–185.
24. Goldberg A, Chavis M, Watkins J, Wilson T. The five-times-sit-to-stand test: validity, reliability and detectable change in older females. Aging Clin Exp Res 2012; 24: 339–344.
25. Goldberg A, Casby A, Wasielewski M. Minimum detectable change for single-leg-stance-time in older adults. Gait Posture 2011; 33: 737–739.