Universidad del País Vasco
Euskal Herriko Unibertsitatea
HEZKUNTZA ETA KIROL FAKULTATEA
FACULTAD DE EDUCACIÓN Y DEPORTE

Máster Universitario en Ciencias de la Actividad Física y del Deporte

TRABAJO FIN DE MÁSTER
CURSO ACADÉMICO 2020-2021

Validity and Reliability of Cardiorespiratory Tests for People with Disabilities: A Systematic Review

Iker Garate Iturzaeta

Vitoria-Gasteiz, 20 de mayo de 2021

Validity and Reliability of Cardiorespiratory Tests for People with Disabilities: A Systematic Review

Trabajo Fin de Máster para optar al Título de Máster en **Ciencias de la Actividad Física y del Deporte**

Presentado por D. Iker Garate Iturzaeta

Tutora Dra. Dª Cristina Granados Domínguez

En Vitoria-Gasteiz, a 20 de mayo de 2021

Firma de/de la alumno/a:

Fdo: D. Iker Garate Iturzaeta

Vº.Bº. del Tutor/a:

Fdo: Dra. Dª Cristina Granados Domínguez

DECLARACIÓN DE FORMATO REVISTA

Formato Sports medicine

https://www.springer.com/journal/40279/submission-guidelines

INDEX

VALIDITY AND RELIABILITY OF CARDIORESPIRATORY TESTS FOR PEOPLE WITH DISABILITIES: A SYSTEMATIC REVIEW

Iker Garate[1*], Cristina Granados[1]

[1] Faculty of Physical Activity and Sports Science, University of the Basque Country, Vitoria-Gasteiz, Spain.

*igarate019@ikasle.ehu.eus

Words: 4204

TITLE: Validity and Reliability of Cardiorespiratory Tests for People with Disabilities: A Systematic Review

ABSTRACT

Background Physical fitness, and especially cardiorespiratory fitness, are very important indicators of health in the general population, and even more so in people with disabilities. In order to measure cardiorespiratory fitness correctly, it is necessary to use valid and reliable tests, and to do so, these psychometric properties must be measured beforehand.

Objective The aim of this systematic review is to know which cardiopulmonary tests have been evaluated for psychometric properties in people with disabilities and which of these tests are reliable and valid for this population.

Methods PubMed, Scopus, SPORTDiscus and Web of Science databases were searched on 30 November 2020. After screening 563 studies, 35 articles met the inclusion criteria: a) participants had any physical, intellectual or sensory disability; b) the test under study measured cardiorespiratory fitness; c) the article provided information on reliability or validity of the test; and d) studies must be original articles. The quality of the studies was assessed according to the COSMIN checklist and this was taken into account when establishing the evidence for each test.

Results Data from a total of 1126 people (370 females and 756 males) have been included in this review, analysing 29 different tests. 23 studies had people with intellectual disabilities as a population, 10 had people with physical disabilities and there were only 2 articles with people with visual disabilities.

Conclusion In people with intellectual disabilities (including those with Down's Syndrome) the six-minute walking test and the shuttle run tests (16 and 20 metre versions) appear to be the tests with the best reliability and validity and the Gold Standard is considered to be an incremental treadmill test. In people with physical disabilities no clear conclusions could be drawn, as the literature seems to be scattered. The incremental wheelchair test has been proposed as a possible Gold Standard versus the arm crank ergometer test for wheelchair-dependent people or wheelchair sports players. In visually impaired people there are very few studies and more research is needed.

KEYPOINTS:

- For people with intellectual disabilities the six-minute walking test seems to be the most valid, although depending on the situation the 20-metre shuttle run test or the 16-metre shuttle run test may also be appropriate.
- In people with physical disabilities more research is needed, but it has been questioned which is the Gold Standard for wheelchair dependent people, with the balance seeming to tip in favour of an incremental test on a wheelchair ergometer versus one performed on an arm crank ergometer.
- Very little research has been done on people with visual impairment and more research is needed.

1 INTRODUCTION

Physical fitness is widely recognised as an indicator of health in both young people [1,2] and adults [3]. One of its most important components is cardiorespiratory fitness, which independently is a major morbidity and mortality predictor in general population [1]. Due to different factors as impairments, activity limitations, or participation restrictions people with disabilities are more likely to be physically inactive [4,5], had poorer cardiorespiratory fitness and they tend to develop more chronic diseases and comorbidities [4–9].

Cardiorespiratory fitness is measured commonly during a laboratory graded maximal exercise test resulting in maximal oxygen consumption (VO_2max) [10], which indicates how much oxygen is the body able to absorb, transport and utilise, but it can be estimated with field tests too [11]. Testing this capacity is important to identify individuals who could benefit from a prevention programme [12] or even for sporting purposes [13–15]. Whatever the objective of the exercise programme, testing is necessary to correctly prescribe, monitor and evaluate it [16,17]. But when it comes to testing, it is not enough to use just any test. The quality of the information obtained will depend on the psychometric properties of the test used, especially reliability and validity, which should be measured beforehand [18].

When assessing the reliability of a test, it must be considered that it has two components: the relative reliability and the absolute reliability [19] (sometimes also called test-retest reliability and measurement error, respectively). The former refers to the extent to which individuals maintain their position in a sample after repeated measurements, while the latter refers to the degree that repeated measurements of individuals vary [19]. The most common statistic used for assess relative reliability is the intraclass correlation coefficient (ICC) or, alternatively, another correlation coefficient. To assess absolute reliability, the standard error of measurement (SEM), the coefficient of variation (CV) or Bland and Altman's 95% limits of agreement (LoA) can be used [19]. Validity describes the degree to which the test reflects what it is intended to measure [20]. This property can be assessed by correlating the test with a Gold Standard (criterion validity, which in turn can be concurrent or predictive), by analysing the correlation between the test results and other measures of the same construct (construct validity, which in turn can be convergent, discriminative or by factor analysis), by contrasting the opinion of experts in the field (content validity) or by using subjective factors (face validity) [20].

Given that cardiopulmonary fitness is, as already mentioned, a very important quality, it is quite possible that the properties of most of the tests used in the literature have already been assessed in the general population. However, in this case, since a very specific population such as people with disabilities is involved, it is necessary to evaluate these tests with a representative sample of that population [21]. It should be noted that for practical reasons, although the authors are aware that disability is multifactorial and not only limited to impairment [5], in this article the word "disability" refers to physical, intellectual and sensory impairments, and so they have been classified as such throughout the document according to the literature.

In the literature, only a few systematic reviews have been conducted on a similar topic, but with different populations or tests than this one [21–23] and only one assessed the quality of the studies [21]. Therefore, as no comparable systematic review exist, the main aim of this review is to know which cardiopulmonary tests have been evaluated for

psychometric properties in people with disabilities and which of these tests are reliable and valid for this population.

2 METHODS

2.1 Literature search

A literature search was conducted on November 30, 2020. Research articles were gathered using PubMed, Scopus, SPORTDiscus and Web of Science database platforms, which represent databases from multiple disciplines related to health and physical activity. No date or language limit was set. The search syntax included the following keywords with relevant Boolean operators inserted: (disability OR disabilities OR disabled) AND cardio* AND test AND (reliability OR validity OR reliability validity).

2.2 Study selection

To be included in this systematic review, studies needed to meet the following criteria: a) participants had any physical, intellectual or sensory disability; b) the test under study measured cardiorespiratory fitness; c) the article provided information on reliability or validity of the test; d) studies must be original articles; and e) full-text article published before December 2020.

The results from the searches were merged, and duplicate records of the same report were removed. Studies were initially screened by the first author on the basis of title and abstract. Relevant abstracts were then selected for a full consideration of the article. Full articles were read in order to ensure the compliance of the inclusion requirements (Figure 1). In case of any concern, the second author was consulted and a consensus was reached.

**** Insert Figure 1 approximately here****

2.3 Data extraction

The first author critically analysed the selected articles and extracted the data of participants (number, sex, age, disability and its characteristics), test studied and statistical analysis on reliability and validity. As in the study selection, if there was any doubt, the second author helped to solve it

2.4 Quality assessment

Articles' methodological quality was evaluated with the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist [24]. It consists of 10 boxes that evaluate different measurement properties as "inadequate", "doubtful", "adequate" and "very good". For this review boxes 6 (relative reliability), 7 (absolute reliability), 8 (criterion validity) and 9 (convergent validity) were rated if applicable. The overall score for each box is determined by the lowest score obtained on any of its items.

2.5 Result analysis and evidence level

For the interpretation of the statistical analysis, ICC values were considered to represent reliability as high (≥0.90), good (0.80-0.89), fair (0.70-0.79) or low (≤ 0.69) [25] and

correlation coefficients, if significant, represented criterion validity as almost perfect (≥ 0.9), very high (0.70-0.89), high (0.5-0.69), moderate (0.3-0.49), small (0.1-0.29), very small (≤ 0.1) [26]. Only studies with "doubtful" or better methodological quality were used for the evidence level analysis. Evidence was considered "strong" if there was one study with "very good" quality or multiple studies with "adequate" quality with similar results; "moderate" if there was one study with "adequate" quality or multiple studies with "doubtful" quality with similar results; and limited if there was only one study with "doubtful" quality [21].

3 RESULTS

3.1 Selected studies

All the selection process is presented in figure 1. The bibliographic search resulted in 707 articles, which after removing duplicates remained at 563 for further analysis. At the end, 35 studies met the inclusion criteria. These papers included a total of 29 different cardiopulmonary tests whose validity (7 cases), reliability (27 cases) or both (21 cases) was reported. 17 of them were field tests and 12 were laboratory tests. All the information on the articles included can be found in table 1.

**** Insert Table 1 approximately here****

3.2 Characteristics of the participants

In the studies sample sizes varied from 4 [27] to 153 [28] and participants' age from 2 years to 80 years old. Most of the studies were carried out with adult people, only one study included people over 65 years of age [29] and nine studies included people under 18 years of age [30–38]. The total number of participants was 1126 (370 females and 756 males).

3.3 Quality assessment and evidence level

Table 2 presents the level of evidence for the validity and reliability of each test based on the results of the best quality articles. Taking into account all the articles and tests, reliability and validity were measured 48 and 29 times respectively. The quality of reliability reports was evaluated as "inadequate" in 23 cases, "doubtful" in 20 cases and "adequate" in 5 cases. The quality of validity reports was "inadequate" in 5 cases, "doubtful" in 16 cases and "adequate" in 7 cases. There was not any "very good" report.

**** Insert Table 2 approximately here****

3.4 Intellectual disabilities

23 of the included studies had as participants people with intellectual disabilities, 984 in total (352 females and 632 males). 8 papers had children or adolescents in the sample [30–35,37,38] and the rest were composed entirely of adults aged 18-65. Only the population of 2 articles was involved in sport [39,40], in the rest of the studies they were sedentary or the level of physical activity was not specified. In most cases their level of intellectual disability was mild or moderate, but in four articles they also included people with severe [30,37,41] or even profound [42] intellectual disabilities. In almost all the

studies the participants had different intellectual disabilities, but the population of five articles consisted exclusively of people with Down syndrome [30,43–46][23,30,44–46]. Except in the study of Yoon et al. [38], where 37% of the population had autism, in the rest of papers the total population was composed of people with intellectual disabilities.

17 different tests (11 field tests and 6 laboratory tests) psychometric properties were studied in this population: the 6-minute walking test (6MWT), 3 shuttle run tests, 2 fixed distance walks, a step test, the Cooper test, 3 fixed distance runs, 4 treadmill tests and 2 submaximal bicycle ergometer tests. The most studied test was the 6-minute walking test which appeared in 11 articles, 12 counting the paper of Ayán-Pérez et al. [43] where they let participants run during the test. In 7 investigations participants performed the test alone [28,30,32,41–43,46], 4 studies used a pacer to assist participants [37,39,44,45] and Temple et al. [40] compared both of them.

In terms of measurement properties, 38 reliability assessments and 20 validity assessments were carried out in people with intellectual disabilities. 13 of them were rated as "inadequate", 20 as "doubtful" and 5 as "adequate" for the first property, and for the second property, 2 were rated as "inadequate", 11 as "doubtful" and 7 as "adequate".

3.5 Physical disabilities

10 of the included papers studied people with physical disabilities, 127 in total (11 females and 116 males). The sample of 1 paper consisted of children [36] and the rest had adults as population, of which only 2 had people over 45 [29,47]. In half of the articles the population was involved in sports [13,16,27,47,48]. The disabilities of the study populations were varied: 4 had cerebral palsy [27,36,49,50], 2 had lower limb amputations [17,29], 1 had spinal cord injuries [47] and the other 3 include people participating in different wheelchair sports (which may include any of aforementioned disabilities): basketball [16], tennis [13] and rugby [48].

12 different tests (6 field tests and 6 laboratory tests) measurement properties were studied in this population: 2 cadence based submaximal wheelchair tests, a shuttle wheelchair test, an intermittent wheelchair test, the 6-minute walking test, a shuttle run test, 3 wheelchair ergometer tests, a bicycle ergometer test, an arm ergometer test and an arm-leg ergometer test. 9 of them were aimed at wheelchair-dependent people, 2 for people who could ambulate independently and 1 for those who could make use of a bicycle ergometer.

In regard to the psychometric properties, 10 reliability assessments and 6 validity assessments were carried out in people with intellectual disabilities. The 10 times reliability was measured it was rated as "inadequate" and validity was evaluated as "inadequate" 2 times and as "doubtful" 4 times.

3.6 Visual disabilities

Only the population of two articles had sensory disabilities, and in both cases, these were visual disabilities. The population of both studies consisted of young adults. In the study of Gulick & Malone [14] the participants were 7 female goalball players and in Silva et al. [15] article they were 8 B1 level football 5-a-side players. The two tests they evaluated were a goalball specific shuttle run test and the 20-metre shuttle run test, respectively. In

both articles they assessed the validity of the tests. The methodology of the first was rated as "doubtful" and while the second was "inadequate".

4 DISCUSSION

The reliability and validity of cardiopulmonary tests have been extensively evaluated, mostly in the non-disabled population [51–54], but not as many systematic reviews have been done in people with disabilities [21–23,55]. Therefore, the aim of the study was to know which cardiopulmonary tests have been evaluated for psychometric properties in people with disabilities and which of these tests are reliable and valid for this population.

Looking at the results of this review, it is clear that the validity and reliability of cardiopulmonary tests has been most investigated in intellectual disabilities [21–23], followed by physical disabilities [55] and finally, with very little research, sensory disabilities [14,15]. Therefore, the existing evidence and the conclusions of this review are in line with this. It is interesting to see the differences between the populations of each type of disability: the proportion of sporty people in the studies is much higher in physical and visual disabilities compared to intellectual disabilities and, roughly, the first two populations are constituted by young adults while in the latter one, children to adults can be found. Therefore, this should be taken into account when interpreting the results and generalising. Moreover, it may be representative of the difference in physical activity habits between these populations and may also give an idea of the different targets of the research.

Regarding psychometric properties, it seems logical that reliability has been investigated more than validity (36 cases vs 28 cases), as it is easier to do the same test twice than to have the possibility to apply the gold standard in each case. This is consistent with previous reviews with similar characteristics [23]. A contribution of this review is that the quality of the included studies has been assessed and this is taken into account when interpreting the results. As in the study by Wouters et al. [21] it can be appreciated that most of the studies have very low ratings. This is because one of the items considered by the COSMIN checklist is the sample size and, due to the design of these studies and the target population, it may be difficult to recruit many people. If this item was not taken into account, the total validity and reliability assessments would change from 28 “inadequate”, 36 “doubtful” and 12 “adequate” to 3 “inadequate”, 17 “doubtful”, 54 “adequate” and 2 “very good”.

4.1 Intellectual disabilities

As mentioned in the results, for people with intellectual disabilities the 6MWT is the most studied test and the one for which there is the most evidence in its favour, followed by the shuttle run test, in its 20-metre and 16-metre versions. This differs from other reviews where they did not find as many studies on the 6MWT and could not draw firm conclusions [21–23], which shows its popularity rise in recent years. Wouters et al. [21] concluded that fixed-distance tests seemed to be the most suitable field tests for children and adolescents with intellectual disabilities, but it has to be said that he assessed different distances as equal (300 yrs - 1mile). In the review by Ayán-Pérez et al. [23] the half-mile run/walk test was found to be the most valid for people with Down's syndrome but with limitations. Only 7 studies with aerobic endurance tests were included in that review. In

agreement with this review Oppewal et al. [22] found that shuttle rung tests, among others, appear to be valid in people with intellectual disabilities. It should be noted that in the reviews by Ayán-Perez et al. [23] and Oppewal et al. [22] the quality of the included studies was not assessed.

The 6MWT has many advantages: it is easy to understand [28], which is very important in this population; it does not require any special equipment (a wide corridor, two cones and a stopwatch are enough) [39], so it has a very low economic cost; it is submaximal, so it is safe even for people who may suffer from health issues [39], which is very frequent in this population [28], and very suitable for people with low cardiopulmonary fitness [28]; and as it involves walking, it does not require any special technique and could be considered specific for all ambulant people, especially elderly people and intellectual disability [41].

On the other hand, one study mentioned that in shuttle run tests it can be difficult for people with Down Sindrome to keep the right paces, because they disoriented when turning around [43]; they require more time and space [28]; an audio pacing device is needed; and being maximal, may be harder and not very suitable for people with reduced cardiorespiratory capacity [28]. Even so, they may have even better validity and reliability values than the 6MWT (table 2) for measuring aerobic endurance and for someone whose usual activity is running may be very suitable. In addition, there is the possibility of using the 16-metre version, instead of the original 20-metre version, for those with lower cardiopulmonary capacity so that the duration of the test is adequate and the fatigue is aerobic [33,34,44,45].

Some studies have shown that 6MWT performance may be related to level of disability [30,41,42], walking economy [41] or lower limb strength [32]. The lower the disability, the better the economy and the higher the lower limb strength, the better the performance [30,32,41,42]. However, there are also studies in which no relationship has been found between 6MWT performance and leg strength [39]. This is probably due to the fact that in less fit people the limiting factor is the ability to use oxygen in the muscles to produce energy and not the central factors. Vis et al. [42] also found that the 6MWT is not suitable for detecting heart diseases. In this population there also appear to be differences between the sexes with men covering longer distances than women [28], but as most of the samples are mixed, the test can be considered valid for both sexes.

Seeing that some studies included in this review tested with pacer and others did not, the research by Temple et al. [40] is interesting. They compared doing the test with or without pacer and concluded that the two methods were similar, although it should be noted that they did not specify the level of disability of the participants, only mentioning that they participated in the Special Olympics and had low support needs. Moreover, it has been suggested that, although the test has a good reliability from the outset, there may be a learning effect, as reliability improves with further attempts [30,41]. Guerra-Balic et al. [41] recommend one full attempt for people with intellectual disabilities, Solway et al. [56] consider two attempts appropriate for people with Down's Syndrome, and Temple et al. [40] considered it sufficient for Special Olympics participants to see someone perform the full test once and do a few laps of the course.

Apart from this, it seems to be clear that people with Down's Syndrome have a lower cardiorespiratory fitness level and that they should therefore be treated as a separate population when it comes to studying them [43,45]. In their case, the best tests also seem

to be the 6MWT and the SRT but in its 16 metre version [43–45]. Boer & Moss [44] recommended testing them one by one with a researcher running alongside and encouraging them continuously and in a standardised way. In general, the literature seems to agree in considering a maximal progressive treadmill test as the gold standard for measuring the maximal oxygen consumption in people with intellectual disabilities [45], given that in all the articles where the criterion validity was assessed, this type of test is used except in one in which a cycloergometer was used [32].

4.2 Physical disabilities

First of all, it should be underlined that, as mentioned above, most of the participants are young adults, and furthermore, 91% of the participants are male, which should be taken into account when making interpretations. Considering the articles on physical disabilities included in this review, it seems that the literature is not going in the same direction, since in 10 articles 12 different tests have been evaluated. As mentioned in the results, 3 tests for people with cerebral palsy who can walk or ride a bicycle can be differentiated from the rest of the tests that are prepared for wheelchair-dependent people or people who practice wheelchair sports, but the quality of all three studies was considered "inadequate" because of their small sample sizes. The result of the 6MWT seems to have a high correlation with walking ability [50], which is in line with what was mentioned in the section on intellectual disabilities, but in the study it was not related to cardiorespiratory fitness comparing it to a gold standard, only to the oxygen consumed during the same test. The shuttle run test does not seem to have a very good absolute reliability [36] and the reliability of the cycloergometer test with only 4 participants did not have a significant correlation between the test and the retest [27].

There is little evidence as all the samples were very small, only for 4 of these tests there is limited evidence (table 2), where the maximal wheelchair ergometer test based on increasing resistance [47] seems to be the most valid, so no solid conclusions can be drawn. Although there seems to be a consensus in the included studies to consider the maximum incremental test on arm crank ergometer as the gold standard for wheelchair dependent people, the study of Morgan et al. [47] questions it for its lack of specificity and mentions that using the wheelchair ergometer based on increasing resistance in their study might be more appropriate, for its greater efficiency and maximal pulmonary ventilation at peak workloads. This is in line with other reviews on the subject which also consider the use of a wheelchair ergometer to be more appropriate, because there may be a difference in the measurement of maximal oxygen consumption, given that less muscle mass is involved in an arm crank ergometer [55,57]. In support of this Bhambhani et al. [27] also concluded that subjects should be tested on their primary mode of ambulation. Morgan et al. [47] acknowledges that it would be better if the cycloergometer used a functional output of participants' work rather than using power as an input.

Of the remaining tests, only the arm crank ergometer test used in the article by Christensen et al. [17] seems to have positive results showing a very high reliability. In that article it is mentioned that the appropriate time for these tests is 5-9 minutes as opposed to the 8-12 minutes recommended for tests using the lower limbs, because of the risk of peripheral fatigue due to the lower muscle mass used. It also notes that there may be a learning effect and therefore strongly recommends prior familiarisation. The intermittent wheelchair test used in Kelly et al. [48] appears to have high relative reliability, but the absolute reliability is poor. In the shuttle wheelchair test it seems that the limiting factor is not aerobic fitness, but the ability to make changes of direction [13],

so modifications could be made to adjust this. For the rest of the tests, no arguments are given to justify why they perform poorly.

4.3 Visual disabilities

Regarding visual impairment, as already mentioned, there is little evidence. It seems that the beep test may be valid [14], but it contains very specific goalball actions, so in principle it would not be generalisable. Moreover, as it only includes 7 females, the conclusions that can be drawn are very limited. In the case of the 20MSRT, in addition to including only 8 males, very low validity results were obtained [15]. Furthermore, the methodology used was inadequate, as they used the ICC to compare two different tests instead of a correlation. Therefore, more research is needed in this area.

5 CONCLUSIONS

This review provides an overview of the validity and reliability of different tests for people with disabilities. In people with intellectual disabilities there is strong evidence in favour of the 6MWT in particular, but in certain cases shuttle run tests (20m or 16m) are also suitable, as they are reliable and valid. The gold standard in this population seems to be clearly an incremental treadmill test, although different protocols are used. In people with physical disabilities more research is needed, as very different tests are used and it is not even clear what the gold standard is for wheelchair-dependent people. It seems that a maximal test performed on a wheelchair ergometer may be more appropriate than one performed on a crank ergometer. In the case of visually impaired people, very little research exists and more research is needed.

6 LIMITATIONS

Due to the search strategy, there may be studies that have been left out of this review. The use of words such as “wheelchair”, “spinal cord injury”, “cerebral palsy”, “endurance” or “aerobic capacity” would probably have resulted in more studies meeting the inclusion criteria.

7 DECLARATIONS

Funding No sources of funding were used to assist in the preparation of this article.

Conflicts of interest Iker Garate and Cristina Granados declare that they have no conflicts of interest relevant to the content of this review.

8 REFERENCES

1. Ortega FB, Ruiz JR, Castillo MJ, et al. Physical fitness in childhood and adolescence: A powerful marker of health. *International Journal of Obesity* 2008;32:1–11.
2. Bermejo-Cantarero A, Álvarez-Bueno C, Martinez-Vizcaino V, et al. Association between physical activity, sedentary behavior, and fitness with health related quality of life in healthy children and adolescents: A protocol for a systematic review and meta-analysis. *Medicine (United States)* 2017;96.
3. Orland Y, Beeri MS, Levy S, et al. Physical fitness mediates the association between age and cognition in healthy adults. Aging Clin. Exp. Res. 2020;33.
4. Carroll DD, Courtney-Long EA, Stevens AC, et al. Vital signs: disability and physical activity--United States, 2009-2012. MMWR. Morb. Mortal. Wkly. Rep. 2014;63:407–13.
5. World Health Organization. *WORLD REPORT ON DISABILITY*. (2011).
6. Wezenberg D, Van Der Woude LH, Faber WX, et al. Relation between Aerobic Capacity and Walking Ability in Older Adults with a Lower-Limb Amputation. Arch. Phys. Med. Rehabil. 2013;94:1714–1720.
7. Remes L, Isoaho R, Vahlberg T, et al. Major lower extremity amputation in elderly patients with peripheral arterial disease: Incidence and survival rates. Aging Clin. Exp. Res. 2008;20:385–393.
8. Hilgenkamp TIM, Reis D, van Wijck R, et al. Physical activity levels in older adults with intellectual disabilities are extremely low. Res. Dev. Disabil. 2012;33:477–483.
9. Hartman E, Smith J, Westendorp M, et al. Development of physical fitness in children with intellectual disabilities. J. Intellect. Disabil. Res. 2015;59:439–449.
10. Mossberg KA, Amonette WE, Masel BE. Endurance training and cardiorespiratory conditioning after traumatic brain injury. Bone 2011;23:1–7.
11. Mayorga-Vega D, Bocanegra-Parrilla R, Ornelas M et al. Criterion-related validity of the distance- and time-based walk/run field tests for estimating cardiorespiratory fitness: A systematic review and meta-analysis. PLoS One 2016;11:1–24.
12. Ruiz JR, Castro-Piñero J, Artero EG, et al. Predictive validity of health-related fitness in youth: A systematic review. Br. J. Sports Med. 2009;43:909–923.
13. de Groot S, Valent LJ, Fickert R, et al. An incremental shuttle wheel test for wheelchair tennis players. Int. J. Sports Physiol. Perform. 2016;11:1111–1114.
14. Gulick DT, Malone LA. Field test for measuring aerobic capacity in paralympic goalball athletes. Int. J. Athl. Ther. Train. 2011;16:22–25.
15. Silva P., Mainenti M., Felicio L., et al. Cardiorespiratory Fitness of Visually Impaired Footballers through Direct and Indirect Methods: A Pilot Study. J. Exerc. Physiol. Online 2005;8:11–25.
16. Laskin J, Slivka DR. A cadence based sub-maximal field test for the prediction of peak oxygen consumption in elite wheelchair basketball athletes. J. Exerc. Physiol. Online 2004;7:8–18.
17. Christensen J, Tang L, Doherty P, et al. Test-retest reliability of a maximal arm cycle exercise test for younger individuals with traumatic lower limb amputations. Eur. J. Physiother. 2020;22:115–120.
18. Currell K, Jeukendrup AE. Validity, reliability and sensitivity of measures of sporting performance. *Sports Medicine* 2008;38:297–316.
19. Bruton A, Conway JH, Holgate ST. Reliability: What is it, and how is it

measured? Physiotherapy 2000;86:94–99.
20. Slater LM, Hillier SL, Civetta LR. The Clinimetric Properties of Performance-Based Gross Motor Tests Used for Children With Developmental Coordination Disorder: A Systematic Review. Pediatr. Phys. Ther. 2010;22:170–179.
21. Wouters M, Evenhuis HM, Hilgenkamp TIM. Systematic review of field-based physical fitness tests for children and adolescents with intellectual disabilities. Res. Dev. Disabil. 2017;61:77–94.
22. Oppewal A, Hilgenkamp TIM, van Wijck R, et al. Cardiorespiratory fitness in individuals with intellectual disabilities-A review. Res. Dev. Disabil. 2013;34:3301–3316.
23. Ayán Pérez C, Martínez-Lemos I, Lago-Ballesteros J, et al. Reliability and Validity of Physical Fitness Field-Based Tests in Down Syndrome: A Systematic Review. J. Policy Pract. Intellect. Disabil. 2016;13:142–156.
24. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. Qual. Life Res. 2012;21:651–657.
25. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability.1. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979, 86:420–8. Psychol. Bull. 1979;86:420–8.
26. Hopkins WG, Marshall SW, Batterham AM, et al. Progressive statistics for studies in sports medicine and exercise science. Med. Sci. Sports Exerc. 2009;41:3–12.
27. Bhambhani YN, Holland LJ, Steadward RD. Maximal aerobic power in cerebral palsied wheelchair athletes: Validity and reliability. Arch. Phys. Med. Rehabil. 1992;73:246–252.
28. Alcántara-Cordero FJ, Gómez-Píriz PT, Sánchez-López AM, et al. Feasibility and reliability of a physical fitness tests battery for adults with intellectual disabilities: The SAMU DIS-FIT battery. Disabil. Health J. 2020;13.
29. Simmelink EK, Wempe JB, Geertzen JHB, et al. Feasibility, safety, and reliability of exercise testing using the combined arm-leg (Cruiser) ergometer in subjects with a lower limb amputation. PLoS One 2018;13:1–14.
30. Casey AF, Wang X, Osterling K. Test-retest reliability of the 6-minute walk test in individuals with down syndrome. Arch. Phys. Med. Rehabil. 2012;93:2068–2074.
31. Beets MW, Pitetti KH, Fernhall B. Peak Heart Rates in Youth With Mental Retardation: Pacer Vs. Treadmill. Pediatr. Exerc. Sci. 2005;17:51.
32. Elmahgoub SS, Van De Velde A, Peersman W, et al. Reproducibility, validity and predictors of six-minute walk test in overweight and obese adolescents with intellectual disability. Disabil. Rehabil. 2012;34:846–851.
33. Fernhall B, Pitetti KH, Vukovich MD, et al. Validation of cardiovascular fitness field tests in children with mental retardation. Am. J. Ment. Retard. 1998;102:602–612.
34. Gillespie M. Reliability of the 20-Metre Shuttle Run for Children With Intellectual Disabilities. Eur. J. Adapt. Phys. Act. 2009;2:7–13.
35. Teo-Koh SM, McCubbin JA. Relationship between peak VO2 and 1-mile walk test performance of adolescent males with mental retardation. Pediatr. Exerc. Sci. 1999;11:144–157.
36. Verschuren O, Bosma L, Takken T. Reliability of a shuttle run test for children with cerebral palsy who are classified at Gross Motor Function Classification System level III. Dev. Med. Child Neurol. 2011;53:470–472.

37. Wouters M, Van Der Zanden AM, Evenhuis HM, et al. Feasibility and reliability of tests measuring health-related physical fitness in children with moderate to severe levels of intellectual disability. Am. J. Intellect. Dev. Disabil. 2017;122:422–438.
38. Yoon TH, Mun YK, Lee JS, et al. Analysis for reliability and validity of gross motor function and health fitness tests for children with developmental disabilities. J. Exerc. Rehabil. 2019;15:667–675.
39. Nasuti G, Stuart-Hill L, Temple VA. The Six-Minute Walk Test for adults with intellectual disability: a study of validity and reliability. J. Intellect. Dev. Disabil. 2013;38:31–38.
40. Temple VA, Alston KF, Elder JJ, et al. The effect of a pacer versus no-pacer on submaximal fitness test results among Special Olympics athletes. Eur. J. Adapt. Phys. Act. 2019;12:6–13.
41. Guerra-Balic M, Oviedo GR, Javierre C, et al. Reliability and validity of the 6-min walk test in adults and seniors with intellectual disabilities. Res. Dev. Disabil. 2015;47:144–153.
42. Vis JC, Thoonsen H, Duffels MG, et al. Six-Minute Walk Test in Patients With Down Syndrome: Validity and Reproducibility. Arch. Phys. Med. Rehabil. 2009;90:1423–1427.
43. Ayan-Perez C, Martínez-Lemos RI, Cancela-Carranl JM. Reliability and convergent validity of the 6-min run test in young adults with Down syndrome. Disabil. Health J. 2017;10:105–113.
44. 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.
45. Boer PH, Moss SJ. Validity of the 16-metre PACER and six-minute walk test in adults with Down syndrome. Disabil. Rehabil. 2016;38:2575–2583.
46. Cabeza-Ruiz R, Alcántara-Cordero FJ, Ruiz-Gavilán I, et al. Feasibility and reliability of a physical fitness test battery in individuals with down syndrome. Int. J. Environ. Res. Public Health 2019;16.
47. Morgan KA, Taylor KL, Tucker SM, et al. Exercise testing protocol using a roller system for manual wheelchair users with spinal cord injury. J. Spinal Cord Med. 2019;42:288–297.
48. Kelly VG, Chen KK, Oyewale M. Reliability of the 30-15 intermittent fitness test for elite wheelchair rugby players. Sci. Med. Footb. 2018;2:191–195.
49. Holland LJ, Bhambhani YN, Ferrara MS et al. Reliability of the maximal aerobic power and ventilatory threshold in adults with cerebral palsy. Arch. Phys. Med. Rehabil. 1994;75:687–691.
50. Maltais DB, Robitaille NM, Dumas F, et al. Measuring steady-state oxygen uptake during the 6-min walk test in adults with cerebral palsy: Feasibility and construct validity. Int. J. Rehabil. Res. 2012;35:181–183.
51. Grgic J, Oppici L, Mikulic P, et al. Test–Retest Reliability of the Yo-Yo Test: A Systematic Review. *Sports Medicine* 2019;49:1547–1557.
52. Singh SJ, Puhan MA, Andrianopoulos V, et al. An official systematic review of the European Respiratory Society/American Thoracic Society: Measurement properties of field walking tests in chronic respiratory disease. *European Respiratory Journal* 2014;44:1447–1478.
53. Moore M, Barker K. The validity and reliability of the four square step test in different adult populations: A systematic review. Syst. Rev. 2017;6.
54. Bellet RN, Adams L, Morris NR. The 6-minute walk test in outpatient cardiac

rehabilitation: Validity, reliability and responsiveness-a systematic review. *Physiotherapy (United Kingdom)* 2012;98:277–286.
55. Goosey-Tolfrey VL, Leicht CA. Field-based physiological testing of wheelchair athletes. Sport. Med. 2013;43:77–91.
56. Solway S, Brooks D, Lacasse Y, et al. A qualitative systematic overview of the measurement properties of functional walk tests used in the cardiorespiratory domain. *Chest* 2001;119:256–270.
57. Eerden S, Dekker R, Hettinga FJ. Maximal and submaximal aerobic tests for wheelchair-dependent persons with spinal cord injury: a systematic review to summarize and identify useful applications for clinical rehabilitation. Disabil. Rehabil. 2018;40:25–35.
58. Montgomery DL, Reid G, Koziris LP. Reliability and validity of three fitness tests for adults with mental handicaps. Can. J. Sport Sci. 1992;17:309–315.
59. Rintala P. Validation of a cardiorespiratory fitness test for men with mental retardation. 1990.
60. Kittredge JM, Rimmer JH, Looney M. Validation of the Rockport Fitness Walking Test for adults with mental retardation. Med. Sci. Sport. Exerc. 1994.
61. Cabeza-Ruiz R, Sánchez-López AM, Trigo ME, et al. Feasibility and reliability of the Assessing Levels of Physical Activity health-related fitness test battery in adults with intellectual disabilities. J. Intellect. Disabil. Res. 2020;64:612–628.
62. Cressler M, Lavay B, Giese M. The reliability of four measures of cardiovascular fitness with mentally retarded adults. Adapt. Phys. Act. Q. 1988;5:285–292.
63. Fernhall B, Tymeson GT. Validation of cardiovascular fitness field tests for adults with mental retardation. Adapt. Phys. Act. Q. 1988;5:49–59.

9 TABLES

Table 1 Characteristics and results of included articles

Test	Study	Disability / sport	Participants	Age	Reliability (Relative & absolute)	Quality	Validity (Criterion)	Quality
Intellectual disabilities								
6MWT	Alcántara-Cordero et al., 2020 [28]	Mild/moderate ID	153M	(18-65)	ICC = 0,82 (0,76 - 0,87) SEM = 41,04 m	2		
	Alcántara-Cordero et al., 2020 [28]	Mild/moderate ID	71F	(18-65)	ICC = 0,73 (0,60 - 0,82) SEM = 38,21 m	2		
	Cabeza-Ruiz et al., 2019 [46]	DS Mild/moderate ID	26M 11F	37,57 (21-58)	ICC = 0,77 (0,595 - 0,874) SEM = 42,26 mMixed factorial ANOVA no differences	3		
	Casey et al., 2012 [30]	DS Mild (20)/ moderate (24)/ severe (11) ID	27M 38F	(11-26)	4 trials (T1, T2, T3, T4) T test only no differences between T3-T4 2 way ANOVA differences between the 3 levels of ID ICC = 0,84 - 0,97 (range) SEM = 12-28 m (range)	2		
	Elmahgoub et al., 2012 [32]	Mild/moderate ID + overweight /obese	15M 24F (Rel) 22M 39F (Val)	Rel-16,8 Val-16,9 (14-22)	T test no differences ICC = 0,82 (0,68 - 0,90) SEM = 29,8 m LoA = -88,2/77,9 m	3	r = 0,31 (abs), 0,69 (rel) (BE VO_{2peak})	2
	Guerra-Balic et al., 2015 [41]	Mild (15) /moderate (18) /severe (13) ID	26M 20F	41 ± 11	3 trials (T1, T2, T3) T test differences between T1 and T2/T3 and differences only between mild and severe level of ID. Repeated measures ANOVA no differences in any trial in any ID level ICC = 0,95 (0,88-0,98) - 0,96 (0,93 -0,98) (range) SEM = 19,9 - 26,6 m (range) LoA = -64,7/51,9 m	3	r = 0,65 (TM VO_{2peak})	2

	Vis et al., 2009 [42]	Mild-profound ID, cardiac restriction (29), Eisenmenger syndrome (Rel)	7M 7F (Rel) 53M 28F (val)	32 (19-44)	T test no differences LoA = -77/62 m CV = 11%	3	No valid for examine cardiac restriction in people with DS.	4
$6MWT_{pacer}$	Boer & Moss, 2016 [44]	DS	24M 19F	33,6 ± 8,6 (19-50)	ICC = 0,93 (0,88 - 0,96) SEM = 21,24 m T test no differences LoA ≃ -55/65 m	3		
	Boer & Moss, 2016 [45]	DS	24M 19F	33,6 ± 8,6 (19-50)			r = 0,78 (TM VO_{2peak}) T test no differences	2
	Nasuti et al., 2013 [39]	ID <32% (SIS) Special Olympics	7M 6F	30,4 ± 7,6 (18-44)	ICC = 0,98	4	r = 0,84 (TM VO_{2peak}) R^2 = 0,67	3
	Wouters et al., 2017 [37]	Moderate/severe ID	25M 12F (30 at the end - 81%)	(2-17)	2-4 weeks difference ICC = 0,78 (0,6-0,8) SEM = 33 m LoA ≃ -95/88 m	3		
	Wouters et al., 2017 [37]	Moderate/severe ID	25M 12F (30 at the end - 81%)	(2-17)	1h difference ICC = 0,95 (0,88-0,98) SEM = 16 m LoA ≃ -49/48m	3		
$6MWT_{pacer/no\ pacer}$	Temple et al., 2019 [40]	ID Low support needs Special Olympics	12M 6F	36,6 ± 10,1 (19-58)	ICC = 0,90 (with pacer) ICC = 0,93 (without pacer)	4	Convergent validity (Between them) r = 0,65-0,65-0,81-0,87 ANOVA no differences	4
$6MWT_{can\ run}$	Ayan-Perez et al., 2017 [43]	DS Mild ID	24M 27F	26,2 ± 7,14 (19-47)	ICC = 0,97 (0,95 - 0,98) LoA = -96,13/81,79m	2	Convergent validity (16MSRT) Correlations varied r = 0,65 - 0,77	3
20MSRT	Beets et al., 2005 [31]	Mild ID (3DS, sex not specified)	31M	15,0 ± 3,36	ICC = 0,90 (0,80 - 0,95) T test no differences	3		
	Beets et al., 2005 [31]	Mild ID (3DS, sex not specified)	11F	15,5 ± 3,88	ICC = 0,91 (0,69 - 0,97) T test no differences	4		
	Fernhall et al., 1998 [33]	Mild/moderate ID	22M 12F	14,3 ± 2,34 (10-17)	Repeated measures ANOVA no differences ICC = 0,97	3	r = 0,74 (TM VO_{2peak}) Convergent validity r = 0,94 (16MSRT) r = -0,62 (600yrs)	2
	Gillespie, 2009 [34]	Mild ID	15M 15F	8 ± 0,63 (6,7-9,1)	ICC = 0,53	3		

	Montgomery et al., 1992 [58]	Moderate ID	18 (sex not especified)	26,3 ± 3,2 (20-35)	ICC = 0,90 ANOVA no differences between the 5 trials	4	r = 0,78 (TM VO_{2peak})	3
16-MSRT	Ayan-Perez et al., 2017 [43]	DS Mild ID	24M 27F	26,2 ± 7,14 (19-47)	ICC = 0,85 (0,735 - 0,91)	2	Convergent validity (6MWT) Correlations varied r = 0,65 - 77	3
	Boer & Moss, 2016 [44]	DS	24M 19F	33,6 ± 8,6 (19-50)	ICC = 0,99 (0,98 - 0,99) SEM = 1,54 shuttles T test no differences LoA ≃ -5/4 shuttles	3		
	Boer & Moss, 2016 [45]	DS	24M 19F	33,6 ± 8,6 (19-50)			r = 0,87 (TM VO_{2peak}) LoA = -5,63/5,62 ml/kg/min T test no differences	2
	Fernhall et al., 1998 [33]	Mild/moderate ID	22M 12F	14,3 ± 2,34 (10-17)	Repeated measures ANOVA no differences ICC = 0,96	3	r = 0,77 (TM VO_{2peak}) Convergent validity r = 0,94 (20MSRT) r = -0,64 (600yrs)	2
15MSRT	Yoon et al., 2019 [38]	ID (22) & Autism (13)	35M	10,31 ± 1,25 (9-12)	ICC = 0,80 (0,65-0,90) LoA-1,68/1,11 levels	3		
Rockport (1 mile walk)	Rintala, 1990 [59]	Moderate ID	19M	26 ± 5,6 (18-38)	ICC = 0,97 r = 0,97	4	r = -0,78, -0,81(TM rel VO_{2peak})	3
	Teo-Koh & McCubbin, 1999 [35]	Mild/moderate ID (4DS)	40M (Rel) 24M (Val)	14,13 ± 1,3 (12,17-16,58)	ICC = 0,97 T-test no differences	3	r = -0,76 (TM VO_{2peak})	3
Rockport equations	Kittredge et al., 1994 [60]	Mild (17) /moderate (8) ID	12M 13F	33,3 ± 7,4	(if > 40s difference between trials, test was repeated) ICC = 0,97	4	r = 0,81 (rel), 0,87 (abs) (TMVO_{2peak}) R^2 = 0,66 (rel), 0,76 (abs) SEE = 4,25 ml/kg/min, 0,30 l/min T test equations overestimate $VO2_{peak}$ (TM)	3
2 km walk	Cabeza-Ruiz et al., 2020 [61]	Mild/moderate ID	20M	(20-60)	ICC = 0,67 (0,35 - 0,86) SEM = 2,10 min LoA ≃ -6,6/5 min	4		
	Cabeza-Ruiz et al., 2020 [61]	Mild/moderate ID	8F	(20-60)	ICC = 0,50 (0,14 - 0,87) SEM = 2,54 min LoA ≃ -4/8min	4		
Step test (CSFT)	Cressler et al., 1988 [62]	Mild/moderate ID	15M 2F	35 (25-44)	ICC = 0,95	3		

	Montgomery et al., 1992 [58]	Moderate ID	18 (sex not especified)	26,3 ± 3,2 (20-35)	ICC = 0,97 ANOVA no differences between the 5 trials	4	r = 0,72 (TM VO_{2peak})	3
Cooper	Cressler et al., 1988 [62]	Mild/moderate ID	15M 2F	35 (25-44)	ICC = 0,81	3		
1,5 mile run	Fernhall & Tymeson, 1988 [63]	Mild ID	6M 14F	29,5 ± 5,6			r = -0,88 (TM VO_{2peak}) (n = 15) R^2 = 0,76	3
300 yrs run	Fernhall & Tymeson, 1988 [63]	Mild ID	15 (at least 9F)	29,5 ± 5,6			r = -0,71 (TM VO_{2peak}) R^2 = 0,46	3
600 yrs run	Fernhall et al., 1998 [33]	Mild/moderate ID	22M 12F	14,3 ± 2,34 (10-17)	Repeated measures ANOVA no differences ICC = 0,98	3	r = -0,8 (TM VO_{2peak}) Convergent validity r = -0,64 (16MSRT) r = -0,62 (20MSRT)	2
TM_{other}	Beets et al., 2005 [31]	Mild ID (3DS, sex not specified)	31M	15,0 ± 3,36	ICC = 0,82 (0,65 - 0,91) T test no differences	3		
	Beets et al., 2005 [31]	Mild ID (3DS, sex not specified)	11W	15,5 ± 3,88	ICC = 0,60 (0,03 - 0,87) T test no differences	4		
TM_{Balke}	Cressler et al., 1988 [62]	Mild/moderate ID	15M 2F	35 (25-44)	ICC = 0,93	3		
$TM_{Balke\ modified}$	Teo-Koh & McCubbin, 1999 [35]	Mild/moderate ID (4DS)	24M	14,13 ± 1,3 (12,17-16,58)	ICC = 0,91 T-test no differences	4		
TM_{Bruce}	Montgomery et al., 1992 [58]	Moderate ID	18 (sex not especified)	26,3 ± 3,2 (20-35)	ICC = 0,93 ANOVA no differences between the 5 trials	4		
$BE_{submax1}$	Cressler et al., 1988 [62]	Mild/moderate ID	15M 2F	35 (25-44)	ICC = 0,64	3		
$BE_{submax2}$	Montgomery et al., 1992 [58]	Moderate ID	18 (sex not especified)	26,3 ± 3,2 (20-35)	ICC = 0,93 ANOVA no differences between the 5 trials	4	r = 0,39 (ns) (TM VO_{2peak})	3
Physical disabilities								
$5CST_{60\ pushes/min}$	Laskin & Slivka, 2004 [16]	W Basketball Different disabilities (1-4,5)	24M (16 reli)	26,1 ± 6,6	ICC = 0,50 LoA = -0,83/1,05 l/min	4	r = 0,49 (AE VO_{2peak}) T test no differences	3

$5CST_{80}$ pushes/min	Laskin & Slivka, 2004 [16]	W Basketball Different disabilities (1-4,5)	24M (16 reli)	26,1 ± 6,6	ICC = 0,62 LoA = -0,66/1,06 l/min	4	r = 0,56 (AE VO_{2peak}) T test no differences	3
SWT	de Groot et al., 2016 [13]	W Tennis Different disabilities	15M				Convergent validity r = 0,40 (ns)(Abs), 0,47 (ns) (Rel) (Gas analyser VO_{2peak})	4
30-15IFT	Kelly et al., 2018 [48]	W Rugby Different disabilities (0,5-3,5)	10M	31,8 ± 7,3 (20-44)	ICC = 0,99 SEM = 1.02 km/h CV = 1,9 % LoA = −0,51/0,61 km/h	4		
6MWT	Maltais et al., 2012 [50]	CP Walk without support	15 (sex not especified)	(20-45)			Convergent validity r = -0.57 (walking ability), - 0.66 (Gas analyser net VO_2)	4
SRT3	Verschuren et al., 2011 [36]	CP GMFCS level 3	8M 5F	12 ± 3	ICC = 0,98 (0,93-0,99) SEM = 0,48 levels LoA ≃ -1,8/2,3 levels	4		
WE_{speed1}	Bhambhani et al., 1992 [27]	Class 3-4 CP_IRSA athletes	6M	24,8 ± 3,7 (19-29)	r = 0,89	4	r < 0,31 (ns) (any BE trial)	3
$WE_{resistance}$	Morgan et al., 2019 [47]	W Spinal cord injury C5-6 to T8-11 ASIA A/B/C 70% sport	10M	33 ± 19,6 (18-60)	ICC = 0,82 (VO_2) ICC = 0,97(P) LoA = -6/4,5 ml/kg/min	4	r = 0,79 (AE VO_2) r = 0,77 (AE P) T test no differences LoA = -4,1/3,6 ml/kg/min	3
WE_{speed2} & BE_{max}	Holland et al., 1994 [49]	Class 3-4 (WE) & Class 5-7 (BE) CP_IRSA	4M 1 F & 2M 2F (only 7 at the end)	25,2 ± 4,7 (22 - 33)	Both test reliability together r = 0,79 (rel), 0,83 (abs)	4		
BE_{max}	Bhambhani et al., 1992 [27]	Class 3-4 CP_IRSA	4M	24,8 ± 3,7 (19-29)	r = 0,92 (ns)	4		
AE	Christensen et al., 2020 [17]	Lower limb amputation Crus-level or above	8M	32.5 ± 4.57 (18-40)	ICC = 0,51 (0,11-0,85)(Vo_{2peak}), 0,74 (0,40-0,93) (T), 0,73 (0,40-0,93) (P) SEM = 0,18 l/min, 0,25 min, 8,16 W LoA = ± 0,53 l/min (mean diff = -0,13 l/min) CV = 14,48 %	4		
ALE	Simmelink et al., 2018 [29]	Unilateral lower limb amputation	14M 3W	54,5 ± 18,6 (25-80)	ICC = 0,84 (0,61-0,94) (VO_{2peak}) & 0,91 (0,77-0,97) (P_{peak}) LoA = -0,56/0,60 l/min	4		

Visual disabilities						
Beep test	Gulick & Malone, 2011 [14]	Goalball	7F	22,7 ± 6,6	r = 0,77 (BE)	3
20MSRT	Silva et al., 2005 [15]	Football B1 level	8M	25 ± 5,3 (17-30)	ICC = 0,58 (TM VO_{2peak}) Wilcoxon underestimate VO_{2peak} vs TM LoA ≃ -3/14 ml/kg/min	4

6MWT = 6 minut walking test; 20/16/15SRT = 20/16/15 metre shuttle run test; TM = Treadmill; CSFT = Canadian standardized fitness test BE = Bicycle ergometer; WE = Wheelchair ergometer; AE = Arm ergometer, ALE = Arm-leg ergometer; 5CST = 5 minutes cadence-based submaximal test; SWT = Shuttle wheelchair test; 30-15IFT = 30-15 intermittent fitness test; SRT3 = GMFCS level III-specific shuttle run test; ID = Intellectual disability; DS = Sown Syndrome; SIS = Supports intensity scale; CP = Cerebral palsy; CP-IRSA = Cerebral Palsy-International Sports and Recreation Association; W = Wheelchair; GMFCS = Gross motor function classification system; M = male; F = Female; ICC = Intraclass correlation coefficient; SEM = Standard error measurement; ANOVA = Analysis of variants; CV = Coefficient of variation; LoA = Limits of agreement; r = correlation coefficient; R^2 = Explained variance; SEE = Standard error of the estimate; VO_2 = Oxygen consumption; P = Power; T = Time.
Some tests varied between studies; if there were large changes, the difference was specified or the variants were simply listed.
If not specified as not significant (ns), it means that the p value is lower than 0,05.
Quality assessment: 1 = very good, 2 = adequate, 3 = doubtful, 4= inadequate.

Table 2 Evidence of validity and reliability of tests

Test	Reliability	Validity
Intellectual disabilities		
6MWT	$^{+++}$ Good	$^{+++}$ High
20MSRT	$^{++}$ High	$^{++}$ Very high
16MSRT	$^{++}$ Good	$^{+++}$ Very high
15MSRT	$^{+}$ Good	N/A
Rockport	$^{+}$ High*	$^{++}$ Very high
2 km walk	?	N/A
Step test	$^{+}$ High*	$^{+}$ Very high
Cooper	$^{+}$ Good*	N/A
1,5 mile run	N/A	$^{+}$ Very high
300 yrs run	N/A	$^{+}$ Very high
600 yrs run	N/A	$^{++}$ Very high
TM_{other}	$^{+}$ Good*	N/A
TM_{Balke}	$^{+}$ High*	N/A
$BE_{submax1}$	$^{+}$ Low*	N/A
Physical disabilities		
$5CST_{60\ pushes/min}$	?	$^{+}$ High
$5CST_{80\ pushes/min}$	?	$^{+}$ Moderate
WE_{speed1}	?	$^{+}$ Small (ns)
$WE_{resistance}$	?	$^{+}$ Very high
Visual disabilities		
Beep test	N/A	$^{+}$ Very high

+++ = Strong evidence level; ++ = Moderate evidence level; + = Limited evidence level; N/A = Not avaliable; ? = indadequate methodology; 6MWT = 6 minut walking test; 20/16/15SRT = 20/16/15 metre run test; TM = Treadmill; BE = Bicycle ergometer; WE = Wheelchair ergometer; 5CST = 5 minutes cadence-based submaximal test; (ns) = not significant correlation at 0,05 level
*Absolute reliability not assessed

10 FIGURES

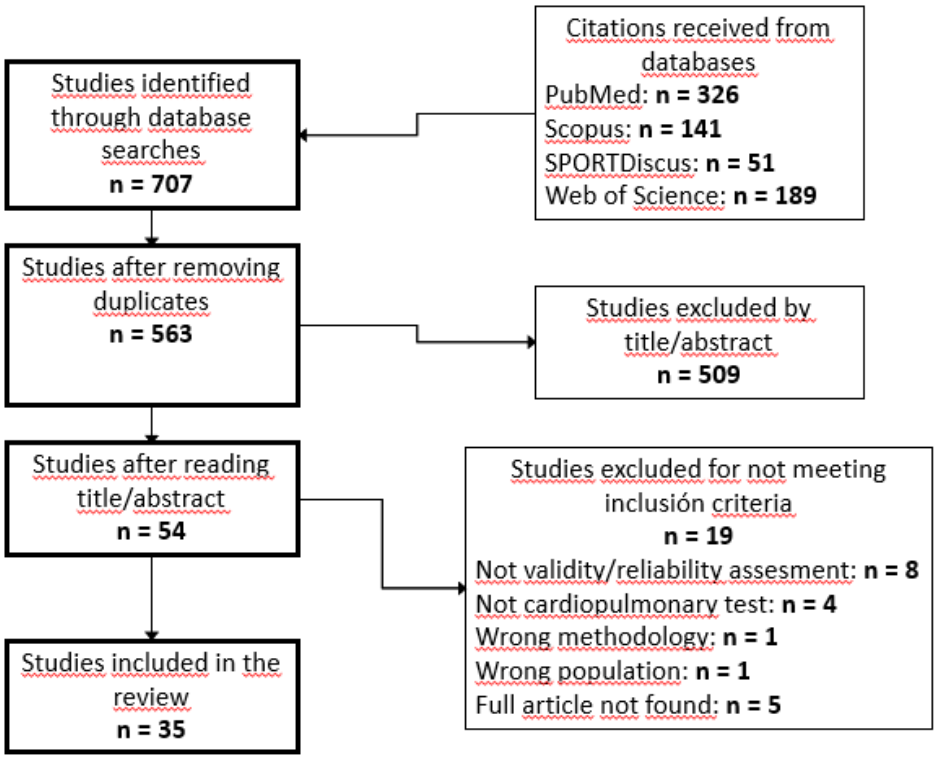

Fig. 1 Study selection flow-chart