Interrater reliability of the new sport-specific evidence-based classification system for Para Va’a

PLEASE CITE THE PUBLISHED VERSION

https://doi.org/10.1123/apaq.2019-0141

PUBLISHER

Human Kinetics

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

Accepted author manuscript version reprinted, by permission, from Adapted Physical Activity Quarterly, 2020, 37 (3): 241-252, https://doi.org/10.1123/apaq.2019-0141. © Human Kinetics, Inc.

Inter-rater reliability of the new sport-specific evidence-based classification system for Para Va'a

Johanna S. Rosén[1], Victoria L. Goosey-Tolfrey[2], Keith Tolfrey[2], Anton Arndt[1,3] & Anna Bjerkefors[1,4]

[1]The Swedish School of Sport and Health Sciences (GIH), Stockholm, Sweden, [2]Peter Harrison Centre for Disability Sport, School of Sport, Exercise and Health Sciences, Loughborough University, UK, [3]Department of Clinical Sciences, Intervention and Technology (CLINTEC), Karolinska Institute, Stockholm, Sweden,[4]Department of Neuroscience, Karolinska Institute, Stockholm, Sweden.

Corresponding Author: Johanna S. Rosén, The Swedish School of Sport and Health Sciences (GIH), Box 5626, SE – 114 86 Stockholm, Sweden. E-Mail: johanna.rosen@gih.se

Co-authors' contact details: Victoria Goosey-Tolfrey V.L.Tolfrey@lboro.ac.uk and Keith Tolfrey K.Tolfrey@lboro.ac.uk are with The Peter Harrison Centre for Disability Sport, School of Sport, Exercise and Health Sciences, Loughborough University, Epinal Way, Loughborough, Leicestershire, LE11 3TU, United Kingdom

Anton Arndt toni.arndt@gih.se and Anna Bjerkefors anna.bjerkefors@gih.se are with the The Swedish School of Sport and Health Sciences (GIH), Box 5626, SE – 114 86 Stockholm, Sweden.

Running title: Reliability of the Para Va'a classification system

Abstract

The purpose of this study was to examine the inter-rater reliability (IRR) of a new evidence-based classification system for Para Va'a. Twelve Para Va'a athletes were classified by three classifier teams consisting of a medical and technical classifier each. IRR was assessed by calculating intra-class correlation for the overall class allocation and total scores of trunk, leg and on-water test batteries and by calculating Fleiss kappa and percentage of total agreement in the individual tests of each test battery. All classifier teams agreed with the overall class allocation of all athletes and all three test batteries exhibited excellent IRR. At a test level, agreement between classifiers was almost perfect in 14 tests, substantial in four tests, moderate in four tests and fair in one test. The results suggest that a Para Va'a athlete can expect to be allocated to the same class regardless of which classifier team conducts the classification.

Keywords*:* outrigger, paracanoe, paddling, Paralympics, impairment

Word Count: 3901

Introduction

Para Va'a is a canoeing sport performed in a Polynesian outrigger canoe, propelled by a single blade paddle on flat or open water, by athletes with physical impairments. Para Va'a makes its debut as a Paralympic sport at the Tokyo 2020 Paralympic Games after the International Paralympic Committee (IPC) in 2018 approved the sport's new sport-specific evidence-based classification system. The system was created in collaboration with international classifiers from the International Canoe Federation (ICF) and is based upon research undertaken by Rosén et al. (2019). In the Paralympic Para Va'a event athletes with the eligible impairment types of limb deficiency, impaired passive range of motion and impaired muscle power affecting trunk and/or legs will compete on flat water over 200 m. The Para Va'a classification involves a medical and a technical assessment performed by a classifier team including a medical and a technical classifier. The medical assessment consists of a trunk and a leg test battery and the technical assessment consists of an on-water test battery. The summarised results from all test batteries are used to allocate the athletes one of three classes: Va'a level 1 (VL1), Va'a level 2 (VL2) or Va'a level 3 (VL3). Athletes competing in VL1 have the most severe impairment and athletes competing in VL3 have the least severe impairment (https://www.canoeicf.com/classification).

In addition to having evidence-based systems, it is also important that the classification systems are reliable. If classifiers are classifying athletes with similar impairments inconsistently, then the credibility of the classifiers and the classification system becomes flawed. Two sports have examined the inter-rater reliability (IRR) of tests used in their classification: wheelchair rugby and Nordic sit-skiing (Altmann, Groen, van Limbeek, Vanlandewijck, & Keijsers, 2013; Pernot, Lannem, Geers, Ruijters, Bloemendal & Seelen, 2011 respectively). The agreement between the classifiers were substantial for both systems, however, a few athletes were allocated to different classes by different classifiers (Altmann et al., 2013; Pernot et al., 2011).

Most of the Paralympic sports that are currently included in the Paralympic Games use classification tests which require limited equipment, are mainly scored using an ordinal scale and can be used by classifiers all around the world (Beckman, Connick & Tweedy, 2017; Tweedy, Connick & Beckman, 2018). Manual muscle tests (MMT) are therefore commonly used in medical assessments to assess impairments affecting muscle strength (Tweedy, Williams & Bourke, 2010; Beckman et al., 2017). One of the disadvantages of using MMT for classification is the difficulty in achieving acceptable reliability (Tweedy & Vanlandewijck, 2011; Beckman et al., 2017). To improve validity, reliability and utility of MMT in classification, Tweedy et al. (2010) suggested that the MMT should be modified by, for example, only assessing movements that are important for performance in the sport and changing the reference range of movement (ROM) from anatomical to sport-specific.

Out of the three classification test batteries for Para Va'a, the leg test battery has the closest resemblance to MMT (Hislop & Montgomery, 1995). All of the classification test batteries have however been developed by taking into consideration the recommendations from Tweedy et al. (2010). The Para Va'a classification test batteries therefore incorporate several different individual tests which all assess movements that are important for performance in the sport, in sport-specific ROM (Rosén et al., 2019). Being able to flex, bend and rotate the trunk and flex and extend the hip, knee and ankle is related to producing a higher force during Va'a paddling (Rosén et al., 2019). The leg tests therefore examine the athlete's abilities to flex and extend the hip, knee and ankle joints and the tests included in the trunk test assess the athlete's ability to flex, extend, rotate and laterally bend the trunk whilst seated. Additionally, the on-water tests assess the athlete's abilities to rotate and flex the trunk and actively move the legs in their sport-specific set-up during paddling on-water. As opposed to the 0 to 5 scale used in MMT, all tests in the Para Va'a classification test batteries are scored on a 0 to 2 scale.

Although the new Para Va'a classification tests have been developed using the recommendations by Tweedy et al. (2010), the tests and the system are new and their reliability must be examined. The overall aim of the study was therefore to examine the IRR of the new Para Va'a classification system. The specific purposes of the study were to examine the IRR in the: a) overall class allocation, b) total score of the trunk, leg and on-water test batteries and c) individual tests in the trunk, leg and on-water test batteries.

Method

Participants

Six internationally level 5 certified medical (3 females, all registered physiotherapists) and technical (2 males and 1 female, all with canoe coaching experience and/or experience as a canoe athlete) Paracanoe classifiers participated in the study. The six classifiers had 6 ± 3 years of experience with a range of experience of 3 to 9 and 2 to 9 years for the medical and technical classifiers, respectively. Twelve Para Va'a athletes (8 males and 4 females: 35 ± 9 years, 72 ± 15 kg, 1.73 ± 0.13 m) with an international competition experience of 4 ± 2 years, from four different countries also volunteered to participate in the study. The athletes had the eligible impairment types of impaired muscle power (health conditions: spinal cord injury or similar (N=4), transverse myelitis (N=1), polio (N=1) or spina bifida (N=1)), impaired passive range of movement (health condition: osteogenesis imperfecta (N=1)), and limb deficiency (health condition: bilateral leg amputation (N=1), unilateral leg amputation (N=3)). The athletes were recruited by sending an email to all European national federations with registered Paracanoe athletes and by posting information about the study on the ICF webpage. The inclusion criteria for the athletes were that they competed at an international level and had an impairment that deemed them eligible for competing in Para Va'a. Following verbal and written information participants provided written consent and completed a health declaration form. Ethical approval for the study was granted by the Regional Ethical Committee, Stockholm, Sweden.

Classification tests, minimal eligibility and class allocation

The trunk test battery consisted of six trunk tests where the athlete sat on a treatment bench and the classifier assessed each athlete's ability to perform trunk flexion, extension, rotation to the left and right and lateral flexion to the left and right (See supplementary material- Trunk test manual). The leg test battery consisted of 14 leg tests performed with or without resistance; bilateral hip, knee and ankle flexion and extension as well as unilateral leg press with both legs (See supplementary material- Leg test manual). Athletes with amputations did not wear their prosthesis/prostheses during the leg tests. The on-water test battery consisted of three tests, which evaluated each athlete's ability to perform trunk flexion, trunk rotation and leg movement during paddling on-water at maximal intensity (See supplementary material- On-Water test manual). The athletes used their preferred boat type and individual adaptations so that the boat set-up replicated what the athlete normally used during competition. Each test for the trunk, leg and on-water test batteries were scored on a 0 to 2 scale. The criteria for each score and test are defined in the test manuals (See supplementary material).

The minimal eligibility criteria for Para Va'a are to have a: a) loss of 10 points or more on one leg or, b) loss of 11 points or more over two legs on the leg test battery, or c) loss of 5 points or more on the trunk test battery with an additional loss of 8 points or more on the leg test battery. The maximal possible score in the leg test that an athlete can have to be eligible for Para Va´a is therefore 18 points. In Para Va'a classification a conversion factor is applied to the total score of the trunk and on-water test batteries to reach the possible maximal score of 18. The trunk test battery score is therefore multiplied by 1.5 and the on-water test battery score is multiplied by 3. Athletes who score 0 on all three test batteries are allocated to the VL1 class.

Athletes who score 15 or above on the trunk test battery are automatically allocated to the VL3 class. If the athletes have a summarised score from all three test batteries of 28 or above, they are also allocated to the VL3 class. Athletes who have a summarised score from all three test batteries of 1-27 are allocated to the VL2 class. Further information about the classification process for Para Va'a can be found on the ICF webpage (https://www.canoeicf.com/classification).

Data collection procedure

Prior to the data collection detailed manuals for the tests were created together with the classifiers whom were members of the ICF's Paracanoe classification sub-committee. The manuals contained instructions and pictures describing how the classifiers should perform each test and how to position, instruct and score the athletes. The manuals were sent to the participating classifiers four weeks prior to the data collection.

The data collection was conducted over two days in Italy in April 2018. The classification process for the study followed the international Para Va'a classification standards and consisted of the trunk, leg and on-water test batteries previously described and also followed the same minimal eligibility criteria and class allocation definitions. As per standard, the classifiers were divided into three teams each comprised of a medical and technical classifier. The medical classifier was present at the technical classifier's test and vice versa, within the same classifier team. For the purpose of the study, each team classified each athlete once, thus all athletes were classified three times. Each classifier team was blinded to the evaluations and the scores of the other classifier teams. The medical evaluations took approximately 30 minutes per athlete and were conducted during day one and half of day two. After three athletes had been classified, each classifier team had 60 minutes to deliberate the scoring of the tested athletes. The technical evaluations were conducted during the other half of day two. All three classifier teams assessed one athlete at a time. All classification tests were filmed as per international Para Va'a classification standards; the technical classifiers filmed the medical tests and vice versa with a standard video camera (Sony RX100V, Tokyo, Japan). The technical tests also included filming the athlete close-up using an action camera (GoPro Hero5 black, San Mateo, CA, USA) mounted in front of the cockpit on the ama of the Va'a (GoPro Jaws mount, San Mateo, CA, USA). For the purpose of this study, the films could also be used by the research team if discrepancies were observed in class allocation of an athlete. If this was the case the research team could examine whether the different classifiers provided the athletes with the same instructions and whether the athletes performed consistently during all classifications.

Statistics

The statistical analysis was carried out in IBM SPSS statistics 25 (IBM, Armonk, NY, USA). The level of significance was set to $p \leq 0.05$. The IRR for the total score of the leg, trunk and on-water test batteries and for the final class allocation was assessed using a two-way random, absolute agreement, single-measures intraclass correlation coefficient (ICC) to assess the degree that classifiers agreed in their classifications of the athletes. The guideline by Cicchetti (1994) was used for the interpretation of ICC where the level of clinical significance is excellent if ICC is between 0.75 and 1.00. For each individual test in all three test batteries, Fleiss kappa was calculated for each individual score (0, 1 and 2) and for the overall test. The guideline by Landis and Koch (1977) was used for the interpretation of Fleiss kappa where kappa <0.00 corresponds to poor agreement, 0.00 to 0.20 slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement and 0.81 to 1.00 almost perfect agreement. 95 % confidence intervals (CI) are reported with the ICC and Fleiss kappa values. Percentage of total agreement was also calculated for all individual tests for each of the three test batteries. The percentage of total agreement was calculated by dividing the number of athletes that all three classifiers were in agreement on with the total number of athletes (N=12) and multiplying with 100.

Results

Class allocation and total score

All three classifier teams were in agreement in the class allocation of the 12 athletes resulting in an ICC of 1.00 (Table 1). Since there were no discrepancies in class allocation, the films obtained during the classifications were not used for further analyses. The total score for the trunk, leg and on-water test batteries showed excellent IRR with ICC values of 0.91 (95 % CI, 0.78 to 0.97, $p < 0.001$), 0.99 (95 % CI, 0.97 to 1.00, $p < 0.001$) and 0.95 (95 % CI, 0.77 to 0.97, $p < 0.001$), respectively (Table 1).

****Table 1 near here****

Trunk tests

Five of the six trunk tests had an overall Fleiss kappa of substantial to almost perfect agreement (Table 2). Furthermore, four of the six trunk tests had a total agreement of 60% or higher. The lowest overall Fleiss kappa for the trunk tests was in trunk extension, exhibiting a Fleiss kappa value of 0.31, which corresponds to fair agreement (Table 2). It was also in this task the lowest percentage of total agreement was seen (33%). The lowest Fleiss kappa was seen in trunk extension for score 1. This kappa value was, however, not significant ($p > 0.05$) meaning that the agreement between the classifiers is not better than would be expected due to chance.

****Table 2 near here****

Leg tests

The leg tests showed, in general, the highest overall Fleiss kappa values, ranging from 0.55 to 1.00 (Table 3). Furthermore, total agreement ranged from 83% to 100% meaning that the classifiers agreed in 10 out of 12 athletes or more for all of the leg tests. For left ankle plantar and dorsiflexion there were relatively low values of Fleiss kappa corresponding to moderate agreement in contrast to high percentages of total agreement (83%). Furthermore, the Fleiss kappa values for score 2 for the left plantar flexion and score 1 for the left dorsiflexion were not significant.

****Table 3 near here****

On-water tests

For the on-water tests the overall Fleiss kappa ranged from 0.42 to 0.91 (Table 4). The leg movement test had both the highest Fleiss kappa value corresponding to almost perfect agreement and the highest percentage total agreement (92%). The two trunk tests in the on-water test battery had overall Fleiss kappa values corresponding to moderate agreement whilst the Fleiss kappa values for the individual scores ranged from poor to substantial agreement (Landis & Koch, 1977). Score 0 showed non-significant Fleiss kappa values for both the trunk tests. Furthermore, total agreement was the lowest in the two trunk tests, 58% and 50% for the trunk flexion and trunk rotation, respectively.

****Table 4 near here****

Discussion

The aim of this study was to examine the IRR of the new evidence-based classification system for Para Va'a. The purposes were to examine the IRR in: a) the overall class allocation, b) the total score of the trunk, leg and on-water test batteries and c) the individual tests in the trunk, leg and on-water test batteries. As expected, the results showed that all the classifier teams (n = 3) were in agreement with the overall class allocation of the twelve athletes. To reach this outcome, the total scores of the trunk, leg and on-water test batteries all showed excellent reliability. On an individual test level the agreement between classifiers was almost perfect for 14 tests, substantial for four tests, moderate for four tests and fair for one test.

As previously mentioned, two sports have examined the IRR of their classification systems. Altmann et al. (2013) examined the IRR of a revised classification system for trunk impairment in wheelchair rugby during two sessions and Pernot et al. (2011) examined the IRR of a test-table-test used in classification of Nordic sit-skiing athletes. The IRR for the wheelchair rugby classification system was shown to be substantial with an overall Fleiss Kappa of 0.76 in the first session and 0.75 in the second session. Pernot et al. (2011) demonstrated a Spearman rank correlation coefficient of 0.95, which according to Altmann et al. (2013), corresponds to an overall Fleiss Kappa of 0.8. In contrast to our study, which showed an ICC of 1.00 for class allocation, the differences between the classifiers in these studies resulted in differences in class allocation. In the studies by Altmann et al. (2013) and Pernot et al. (2011) one individual test battery formed the basis for class allocation. In the Para Va'a classification system however, class allocation is based upon the results of three test batteries, which all demonstrated excellent reliability. Since three test batteries are used for class allocation, it allows for minor differences between classifiers without it affecting class allocation.

All test batteries demonstrated ICC > 0.90 indicating excellent reliability. The reason for the high IRR in the test batteries may be due to the fact that they were developed by following the previously mentioned recommendations from Tweedy et al. (2010). Even though these recommendations were intended for adaptation of MMT for classification, they also worked well as guidelines for developing classification tests in general. In addition to following these recommendations, changing the scale to a 0 to 2 scale instead of a 0 to 5 scale commonly used in MMT, may also be a reason for the high reliability in Para Va'a classification since the differences between the scores are more distinct. The application of MMT for classification purposes has previously been questioned (Connick, Beckman, Deuble & Tweedy, 2016; Tweedy, Beckman & Connick, 2014). This is primarily due to difficulties in achieving acceptable IRR (Beckman et al., 2017). Our results interestingly showed that the leg test battery, which is the test most closely resembling MMT, had the highest IRR. It has previously been shown that the reliability of MMT increases if the raters are well-trained in the tests and if the test descriptions are good (Escolar et al., 2001). The classifiers in this study were all experienced classifiers and had thorough training in these tests.

The leg test battery did not only have the highest ICC value but also demonstrated in general the highest level of agreement between classifiers on an individual test level. Twelve of the fourteen leg tests exhibited overall Fleiss kappa values of almost perfect agreement. Left ankle plantar- and dorsiflexion however exhibited overall Fleiss kappa values corresponding to moderate agreement. The Fleiss kappa was however accompanied with a high agreement of 83% for both tests. The reason for the discrepancies between the low value of Fleiss kappa and the high percentage of total agreement is possibly due to the low prevalence of athletes scoring 1 and 2 in these two tests. Since the minimal eligibility criteria is to have loss of at least ten points in one leg, athletes usually have an impairment affecting at least the ankle, resulting in the majority of the athletes scoring 0 in the ankle. Low values of Fleiss kappa with high percentages of agreement indicate a skewed distribution of scores, which was apparent for these leg tests. Since kappa statistics are influenced by the prevalence of entities for each score, it may not be the most appropriate statistic to assess reliability for these cases (Feinstein & Cicchetti, 1990).

The trunk extension test in the trunk test battery and the two trunk tests in the on-water test battery also exhibited a lower value of overall Fleiss kappa as well as poor and non-significant Fleiss kappa values for individual scores. Few athletes were given a score of 0 in the trunk tests in the on-water test battery. The percentages of total agreement for these tests were however not as high as for the left plantar- and dorsiflexion tests in the leg test battery. The lower reliability observed in the trunk extension test in the trunk test battery and the two trunk tests in the on-water test battery might be due to difficulties in scoring these movements because the athletes can use different compensation strategies to perform the movement. Athletes with impairment affecting the trunk can compensate during the trunk tests by using a unique muscle activation pattern and new muscle synergies such as using upper trunk muscles with intact innervation or using normally non-postural upper trunk muscles (Seelen, Potten, Drukker, Reulen, & Pons, 1998; Potten, Seelen, Drukker, Reulen, & Drost, 1999; Bjerkefors, Carpenter, Cresswell, & Thorstensson, 2009). Furthermore, athletes with impairments that prevent them from activating muscles surrounding the pelvis, e.g. hip flexor and extensor muscles, might also compensate during the trunk tests with trunk kyphosis or lordosis. Distinguishing the movement caused by using a compensation strategy can be difficult for the classifiers to examine because the compensation might make the movement look exaggerated. The description in the trunk test manual for score 0 for the trunk flexion/extension tests state that the "athlete cannot flex or extend without compensation by kyphosis/lordosis or cannot resume straight position without support". Furthermore, the description for score 1 state that the "athlete may compensate to resume straight position". These unclear descriptions of the compensations might result in difficulties for the classifiers to distinguish the difference between score 1 and 0. This may also explain why the Fleiss kappa values are the lowest for these scores especially for the trunk extension test. The difficulties involved in distinguishing trunk impairment from compensation strategies during classification have previously been discussed by Altmann et al. (2013) and it was suggested that test descriptions should place emphasis on describing these difficulties. Furthermore, it has previously been shown that reliability can increase if test descriptions in classification manuals are made more clear (Altmann et al. 2013). To further increase the reliability of these tests in the Para Va'a classification system the test descriptions should be clarified and the possible usage of compensation strategies should be discussed in the trunk and on-water test manuals. The IPC position stand highlights that valid methods of assessing impairment in classification should be: reliable, objective, ratio-scaled, precise, only measure the specified body structure or function, be as training resistant as possible and parsimonious (Tweedy & Vanlandewijck, 2011). Even though the current Para Va'a classification system shows high reliability, it needs to be revised in the future by creating ratio scaled and more objective classification tests in order to fully follow the IPC position stand (Tweedy & Vanlandewijck, 2011).

The main limitation of this study was that there was a limited number of athletes included. The study was conducted before the season started and it was not conducted during a competition. The athletes therefore had to travel and stay at the data collection location for two days in order to partake in the study. This combined with the stress many athletes experience in partaking in classification made it challenging to recruit athletes.

Conclusion

All classifier teams were in agreement with the overall class allocation of the twelve athletes and all test batteries showed excellent reliability (ICC > 0.90). Even though discrepancies between classifiers were seen on an individual test level, this did not affect the overall class allocation. It can therefore be expected that a Para Va'a athlete will be allocated to the same class regardless of which classifier team conducts the classification in the new evidence-based classification system for Para Va'a.

Declaration of interest

The authors report no conflict of interest.

References

Altmann, V. C., Groen, B. E., van Limbeek, J., Vanlandewijck, Y. C., & Keijsers, N. L. (2013). Reliability of the revised wheelchair rugby trunk impairment classification system. *Spinal Cord*, *51*(12), 913-918. doi: 10.1038/sc.2013.109

Beckman, E.M., Connick, M.J. & Tweedy, S.M. (2017). Assessing muscle strength for the purpose of classification in Paralympic sport: A review and recommendations. Journal of Science and Medicine in Sport, *20*(4), 391-396. doi.org/10.1016/j.jsams.2016.08.010

Bjerkefors, A., Carpenter, M.G., Cresswell, A.G. & Thorstensson, A. (2009). Trunk muscle activation in a person with clinically complete thoracic spinal cord injury. *Journal of Rehabilitation Medicine*, *41*(5), 390-392. doi.org/10.2340/16501977-0336.

Cicchetti, D.V. (1994). Guidelines, criteria and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284-290. doi.org/10.1037/1040-3590.6.4.284

Connick, M.J., Beckman, E.M., Deuble, R. & Tweedy, S.M. (2016). Developing tests of impaired coordination for Paralympic classification: normative values and test-retest reliability. *Sports Engineering*, *19*(3), 147-154. doi.org/10.1007/s12283-016-0199-5

Escolar, D.M., Henricson, E.K., Mayhew, J., Florence, J., Leshner, R., Patel, K.M. & Clemens, P.R. (2001). Clinical evaluator reliability for quantitative and manual muscle testing measures of strength in children. *Muscle Nerve*, *24*(6), 787-793. doi.org/10.1002/mus.1070

Feinstein, A.R. & Cicchetti D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543-549. doi.org/10.1016/0895-4356(90)90158-L

Hislop, H.J. & Montgomery, J. (1995). *Daniels and Worthingham's Muscle Testing – Techniques of manual examination*. 6th ed. Philadelphia: W.B. Saunders Co.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. doi.org/10.2307/2529310

Pernot H.F.M., Lannem, A.M., Geers, R.P.J., Ruijters, E.F.G., Bloemendal, M., & Seelen, H.A.M. (2011). Validity of the test-table-test for Nordic skiing for classification of Paralympic sit-ski sports participants. *Spinal Cord*, *49*(8), 935-941. doi: 10.1038/sc.2011.30

Potten, Y.J., Seelen, H.A., Drukker, J., Reulen, J.P. & Drost, M.R. (1999). Postural muscle responses in the spinal cord injured persons during forward reaching. *Ergonomics*, *42*(9), 1200-1215. doi.org/10.1080/001401399185081

Rosén, J.S., Arndt, A., Goosey-Tolfrey, V.L. Mason, B.S., Hutchinson, M.J., Tarassova, O. & Bjerkefors, A. (2019). The impact of impairment on kinematic and kinetic variables in Va'a paddling: towards a sport-specific evidence-based classification system for Para Va'a. *Journal of Sports Sciences*, *37*(17), 1942-1950. doi.org/10.1080/02640414.2019.1606763

Seelen, H.A., Potten, Y.J., Drukker, J., Reulen, J.P. & Pons, C. (1998). Development of new muscle synergies in postural control in spinal cord injured subjects. Journal of Electromyography and Kinesiology, *8*(1), 23-34. doi.org/10.1016/S1050-6411(97)00002-3

Tweedy, S.M., Beckman, E.M. & Connick, M.J. (2014). Paralympic classification- conceptual basis, current methods and research update. *PM&R*, *6*(8 Suppl), S11-S17. doi:10.1016/j.pmrj.2014.04.013

Tweedy, S.M., Connick, M.J. & Beckman, E.M. (2018). Applying scientific principles to enhance Paralympic classification now and in the future- A research primer for rehabilitation specialists. *Physical Medicine and Rehabilitation Clinics of North America*, *29*(2), 313-332. doi: 10.1016/j.pmr.2018.01.010

Tweedy, S. M., & Vanlandewijck, Y. C. (2011). International Paralympic Committee position stand-background and scientific principles of classification in Paralympic sport. *British Journal of Sports Medicine*, *45*(4), 259-269. doi: 10.1136/bjsm.2009.065060

Tweedy, S. M., Williams, G., & Bourke, J. (2010). Selecting and modifying methods of manual muscle testing for classification in Paralympic sport. *European Journal of Adapted Physical Activity*, *3*(2), 7-16. doi: 10.5507/euj.2010.005

Table 1. Total scores for the trunk, leg and on-water test batteries and the class allocation for twelve Para Va'a athletes classified by three classification teams (R1, R2 and R3). The conversion factors of 1.5 and 3 have been applied to the trunk and on-water test batteries respectively.

	Trunk test			Leg test			On-water test			Class allocation		
Athlete	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**	**R1**	**R2**	**R3**
1	9.0	9.0	7.5	0.0	0.0	0.0	9.0	12.0	6.0	VL2	VL2	VL2
2	0.0	0.0	0.0	0.0	0.0	0.0	6.0	3.0	6.0	VL2	VL2	VL2
3	9.0	10.5	9.0	0.0	2.0	0.0	12.0	9.0	9.0	VL2	VL2	VL2
4	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	VL3	VL3	VL3
5	10.5	9.0	12.0	10.0	10.0	10.0	12.0	9.0	9.0	VL3	VL3	VL3
6	7.5	10.5	6.0	0.0	0.0	0.0	6.0	6.0	6.0	VL2	VL2	VL2
7	3.0	9.0	0.0	0.0	3.0	2.0	3.0	6.0	3.0	VL2	VL2	VL2
8	9.0	12.0	10.5	0.0	0.0	0.0	3.0	6.0	6.0	VL2	VL2	VL2
9	18.0	18.0	18.0	16.0	17.0	18.0	18.0	18.0	18.0	VL3	VL3	VL3
10	16.5	18.0	18.0	14.0	14.0	14.0	15.0	18.0	18.0	VL3	VL3	VL3
11	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	18.0	VL3	VL3	VL3
12	13.5	18.0	13.5	12.0	12.0	8.0	15.0	15.0	9.0	VL3	VL3	VL3

Table 2. Fleiss kappa (95% CI) for scores 0, 1 and 2 and overall Fleiss kappa and percentage of total agreement for the trunk tests.

	0	1	2	Overall	% agreement
Flexion	0.87 (0.54-1.20)	0.61 (0.28-0.93)	0.77 (0.44-1.10)	0.75 (0.52-0.98)	75 %
Extension	0.36 (0.03-0.68)	0.000[NS] (-0.33-0.33)	0.55 (0.22-0.88)	0.31 (0.10-0.54)	33 %
Right rotation	0.72 (0.39-1.05)	0.52 (0.19-0.85)	0.67 (0.34-0.99)	0.62 (0.36-0.87)	58 %
Left rotation	0.72 (0.39-1.05)	0.53 (0.21-0.86)	0.67 (0.34-0.99)	0.62 (0.36-0.88)	67 %
Right side shift	0.77 (0.44-1.10)	0.66 (0.33-0.98)	0.78 (0.45-1.10)	0.73 (0.47-0.98)	75 %
Left side shift	0.77 (0.44-1.10)	0.78 (0.45-1.10)	0.88 (0.55-1.21)	0.82 (0.57-1.10)	83 %

[NS] Not significant ($p > 0.05$)

Table 3. Fleiss kappa (95% CI) for scores 0, 1 and 2 and overall Fleiss kappa and percentage of total agreement for the leg tests.

Joint	Side	Test	0	1	2	Overall	% agreement
Hip	**Right**	Flexion	0.89 (0.56-1.22)	0.82 (0.50-1.15)	1.00 (0.67-1.33)	0.91 (0.67-1.15)	92 %
		Extension	0.78 (0.45-1.10)	0.65 (0.32-0.97)	1.00 (0.67-1.33)	0.82 (0.58-1.10)	83 %
	Left	Flexion	0.89 (0.56-1.22)	0.82 (0.50-1.15)	1.00 (0.67-1.33)	0.91 (0.67-1.15)	92 %
		Extension	0.89 (0.56-1.21)	0.84 (0.51-1.17)	1.00 (0.67-1.33)	0.91 (0.68-1.15)	92 %
Knee	**Right**	Flexion	0.88 (0.55-1.21)	0.82 (0.50-1.15)	1.00 (0.67-1.33)	0.89 (0.65-1.14)	92 %
		Extension	1.00 (0.67-1.33)	1.00 (0.67-1.33)	1.00 (0.67-1.33)	1.00 (0.76-1.24)	100 %
	Left	Flexion	1.00 (0.67-1.33)	1.00 (0.67-1.33)	1.00 (0.67-1.33)	1.00 (0.76-1.24)	100 %
		Extension	1.00 (0.67-1.33)	1.00 (0.67-1.33)	1.00 (0.67-1.33)	1.00 (0.76-1.24)	100 %
Ankle	**Right**	Plantar flexion	1.00 (0.67-1.33)	-	1.00 (0.67-1.33)	1.00 (0.67-1.33)	100 %
		Dorsiflexion	1.00 (0.67-1.33)	-	1.00 (0.67-1.33)	1.00 (0.67-1.33)	100 %
	Left	Plantar flexion	0.77 (0.44-1.10)	0.44 (0.11-0.76)	-0.03^{NS} (-0.36-0.30)	0.55 (0.27-0.83)	83 %
		Dorsiflexion	0.77 (0.44-1.10)	0.27^{NS} (-0.05-0.60)	0.47 (0.14-0.80)	0.55 (0.30-0.81)	83 %
Leg press	**Right**		0.87 (0.54-1.20)	0.77 (0.44-1.10)	1.00 (0.67-1.33)	0.88 (0.64-1.13)	92 %
	Left		1.00 (0.67-1.33)	1.00 (0.67-1.33)	1.00 (0.67-1.33)	1.00 (0.76-1.24)	100 %

^{NS} Not significant (p > 0.05)

Table 4. Fleiss kappa (95% CI) for scores 0, 1 and 2 and overall Fleiss kappa and percentage of total agreement for the on-water tests.

	0	**1**	**2**	**Overall**	**% agreement**
Trunk flexion	-0.03^{NS} (-0.36-0.30)	0.44 (0.11-0.76)	0.55 (0.23-0.88)	0.47 (0.17-0.77)	58 %
Trunk rotation	-0.09^{NS} (-0.42-0.24)	0.33^{NS} (0.00-0.65)	0.67 (0.34-0.99)	0.42 (0.15-0.69)	50 %
Leg movement	1.00 (0.67-1.33)	0.82 (0.50-1.15)	0.87 (0.54-1.20)	0.91 (0.67-1.15)	92 %

^{NS} Not significant ($p > 0.05$)