jamk

# Evaluation of the inter- and intra-rater reliability of the Finnish Gymnastics Federation's Strength Tests: A Test-Retest

Anni Laitinen

Master’s thesis
January 2024
Master's Degree Programme in Sport and Exercise Physiotherapy

jamk | Jyväskylän ammattikorkeakoulu
University of Applied Sciences

jamk

**Description**

**Laitinen, Anni**

**Evaluation of the inter- and intra-rater reliability of the Finnish Gymnastics Federation's Strength Tests: A Test-Retest**

Jyväskylä: Jamk University of Applied Sciences, January 2024, 65 pages

Degree Programme in Sport and Exercise Physiotherapy. Master's thesis.

Permission for open access publication: Yes

Language of publication: English

**Abstract**

Competing gymnasts have a high risk of injury due to the large amount of training hours and the demands of the sport (Campbell et al. 2019; Thomas & Thomas, 2019). In high level gymnasts, the repetitive performance of gymnastic movements is not enough to gain the optimal levels of strength and power needed within the sport (Deley et al. 2010). An increase in strength levels increases performance and gymnastic skills (Sawczyn et al. 2016; Dallas et al. 2021). Testing sport-like strength is important for young gymnasts and measuring muscle strength is an important performance parameter (Osmala et al. 2021; Gasparetto, 2022). There are no current such testing batteries available, therefore there is a need for meaningful, reliable, and sensitive outcome gymnastic-specific fitness tests (Salse-Batán et al. 2022). This test-retest evaluates the inter- and intra-rater reliability of the Finnish Gymnastics Federation's strength tests with a two-day time interval between the testing days. This evaluation is the very first step in building an evidence-based test battery to monitor athletes in any sport. These tests are designed for aesthetic group gymnasts and rhythmic gymnasts with the purpose of measuring gymnasts' sport-specific strength and performance.

Tester one was more reliable between the two testers and the results on the second testing day were more reliable compared to the first testing day. In addition, there was a slight decline in the overall score on the second testing day meaning the athletes performed with higher scores on the first testing day. When analysing the athletes total scores, there can be seen an agreement between the measurements of the two testers, showing the athlete's total scores correlating with each other's. The overall strength tests scores can be used to identify those athletes with lower and higher overall strength. ICC-values for the athlete's total scores varied between testers 0.880–0.855 and between testing days 0.891-0.957 meaning good to excellent reliability based on the 95% confident interval of the ICC estimate.

To conclude, clearer guidelines and testing protocols are recommended and more distinctive differences between the scores are needed. Tests assessing leg strength and power are beneficial for the performance and can be used as a tool for injury prevention in AGG, therefore should be added to the future test battery. Three strength tests out of nine have proven to be reliable and have a high sport specificity. Athlete's total scores of the current strength tests can be used to identify "stronger" and "weaker" athletes based on the reliable results found in this study. An athlete identification tool was created to identify stronger to weaker athletes. This tool is simple to use, easy to implement and is designed to benefit and support the athlete.

**Keywords/tags (subjects)**

Gymnastics, Aesthetic Group Gymnastics, Female Athlete, Muscle Strength, Youth Strength Training, Strength Testing, Test-retest

**Contents**

**Figures**

**Tables**

# 1 Introduction

Gymnastics is an aesthetic sport in which performance is based on sport skills and their mastery. In aesthetic group gymnastics (AGG) adolescent females are the largest population actively training and competing in the sport. However, competing gymnasts have a high risk of injury due to the large amount of training hours and the demands of the sport. Female gymnasts have a high injury incidence of 9.37 per 1000 athlete exposures. Most injuries in female gymnasts are primarily in the lower extremities, but also occur on the upper limbs, back area, and spine. (Pei et al. 2022; Campbell et al. 2019; Thomas & Thomas, 2019)

Young athletes benefit from sufficient performance to cope with the multiple demands of an aesthetic sport. In AGG, sufficient strength characteristics support athlete's performance and prevents injuries, especially among young athletes (Alwasif, 2019; Thomas & Thomas, 2019). Strength exercises should answer to the demands of gymnastics, as well as strength characteristics should be increased and focused on to achieve a good gymnastics technical level (Alwasif, 2019). Mobility and a large range of motion of the joints play an important role in gymnastics. However, without sufficient strength, gymnasts will not be able to take advantage of the existing mobility characteristics as part of the sport performance (Osmala et al. 2021).

Skill alone will not guarantee the level of strength needed in gymnastics (Alwasif, 2019). In comparison, an increase in strength levels increases performance as well as gymnastic skills (Sawczyn et al. 2016; Dallas et al. 2021). A gymnast needs a high level of strength to learn the correct technical movements whereas a low level of strength has a negative effect on the development of technical skills (Dallas et al. 2011). In addition, in high level gymnasts, the repetitive performance of gymnastic movements is not enough to gain the optimal levels of strength and power needed within the sport (Deley et al. 2010). Measuring muscle strength is an important performance parameter and therefore strength should be measured throughout the athlete’s training life cycle (Gasparetto, 2022). Currently there are no validated tools for this purpose with the required validity and reliability (Salse-Batán et al. 2022).

This test-retest evaluates the inter- and intra-rater reliability of the Finnish Gymnastics Federation’s strength tests. These strength tests are designed for AGG and rhythmic gymnastics with the purpose to measure gymnasts’ sport-specific strength and performance. This test battery is a tool

for coaches to monitor and measure strength in training situations and highlight the importance of strength needed in the sport. The strength tests, the guidelines, and the scoring system have been developed by the experts of the Finnish Gymnastics Federation.

# 2 Theoretical basis

The PubMed database was used to attain high quality resources for the theoretical basis of this test-retest study. The search was successful, even though a lot of the material could not be found through the PubMed database. A grey literature search was also performed using search engines such as Google Scholar and ResearchGate to supplement the search. Appendix 1 states the overall search process as well as the search terms, Boolean operators, and the search strategy. The number of studies included can be seen in Figure 1.

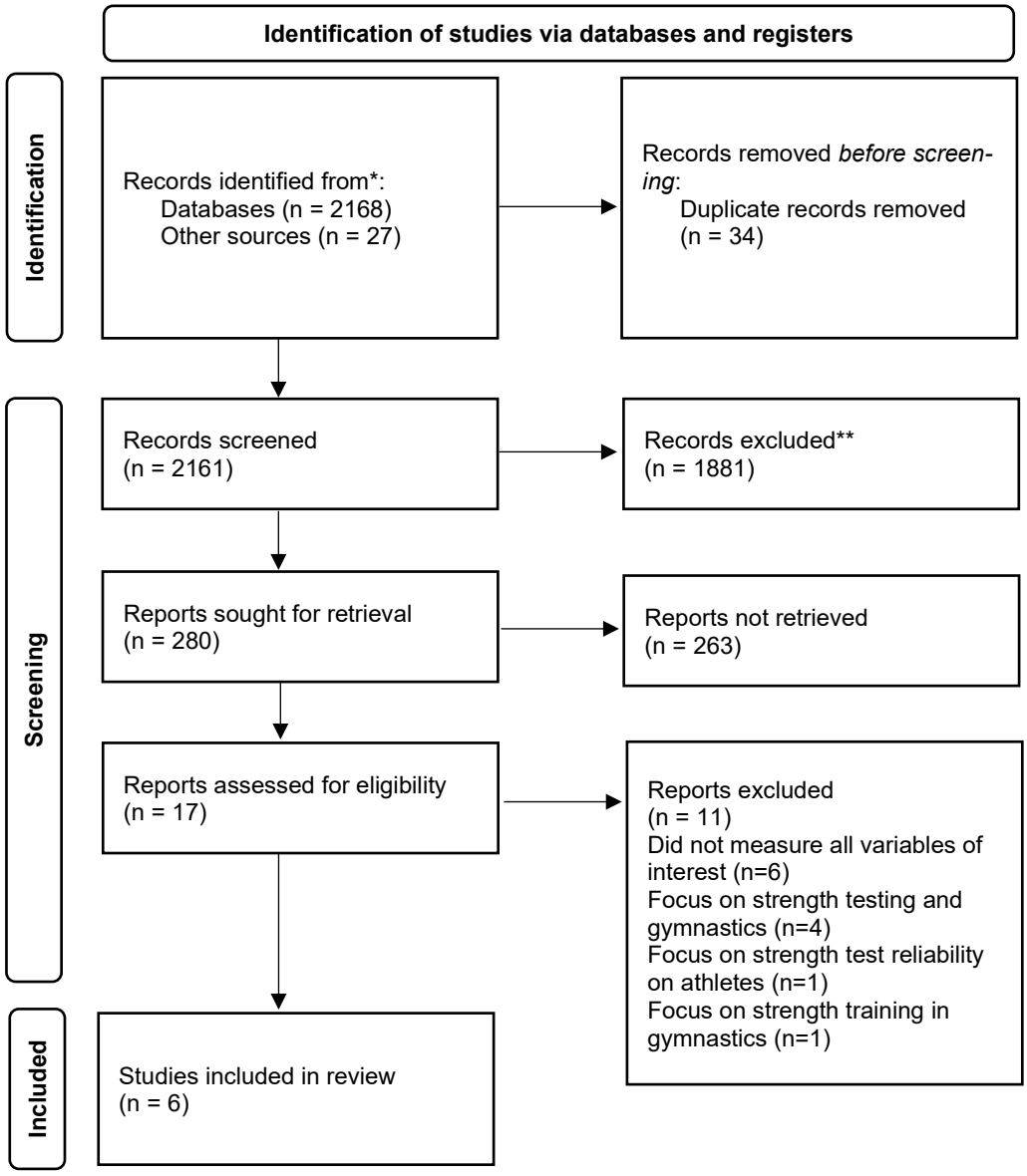



Figure 1. PRISMA Flow chart

## 2.1 Aesthetic group gymnastics

AGG is a high-level competitive sport that involves harmonious, rhythmic, and dynamic movements that naturally flow from one movement to the next. Physical qualities include flexibility, speed, strength, coordination, and the ability to move effortlessly. The versatile and varied body movements include body waves and swings, balances, pivots, jumps and leaps, dance steps as well as lifts. In a choreography, music is interpretated through expression and movements to create a story by a team consisting of 6 to 10 gymnasts. (Santalov, n.d.; Suomen Voimisteluliitto, n.d.)

## 2.2 Muscle strength and youth strength training

Muscle strength has been defined as the ability to exert force on an external object or resistance (Suchomel et al. 2016; Grgic et al. 2020). Overload, specificity, progression, individuality, and reversibility are all principles of strength training (Sands et al. 2012). Resistance training improves athletic performance by increasing muscular strength, power and speed, hypertrophy, local muscular endurance, motor performance, balance, and coordination (Kraemer & Ratamess, 2004).

Strength training in children has proven to be safe and efficient when properly designed and when competently supervised youth resistance training has beneficial effects for health, strength, and athletic performance (Faigenbaum & Micheli, 2017; Behm et al. 2017; Dahab & McCambridge, 2009; Faigenbaum et al. 2009). To the contrary of current beliefs, strength training does not harm the growth of children or damage developing growth plates (Faigenbaum & McFarland, 2016; Lloyd et al. 2014). It is safe for children and adolescents to participate in strength training when they are emotionally mature to follow and accept directions. Strength training has benefitted many seven- and eight-year-old girls and boys (Faigenbaum & Micheli, 2017). In addition, strength training decreases the incidence of sports-related injuries by increasing the strength of tendons, ligaments, and bones (Dahab & McCambridge, 2009; Behm et al. 2017; Faigenbaum & Micheli, 2017). Therefore, neglecting strength training as a part the normal training would be at the harm to the athlete. Currently strength training in AGG is limited and often overlooked, therefore strength training is either being neglected or avoided in this population. One reason for the avoidance of strength training could be the repetitive use of gymnastic movements, which is not enough to gain muscle strength (Deley et al. 2010). Therefore, only focusing on the repetitive per-

formance of the gymnastic movements alone, instead of combining it with strength training, exposes the athlete to more risk (Alwasif, 2019; Thomas & Thomas, 2019; Sawczyn et al. 2016; Dallas et al. 2021).

Youth strength training should still stay safe, effective, and fun. Children are anatomically, physiologically, and psychologically less mature, which is why training philosophies and strength training guidelines designed for adults should not be imposed on children. Strength resistance training protocol for children should involve training 2-3 times a week with moderate loads (50-60% 1RM) and higher repetitions (15-20). Youth resistance training guidelines can be seen in Figure 2. (Faigenbaum & Micheli, 2017; Faigenbaum & McFarland, 2016)

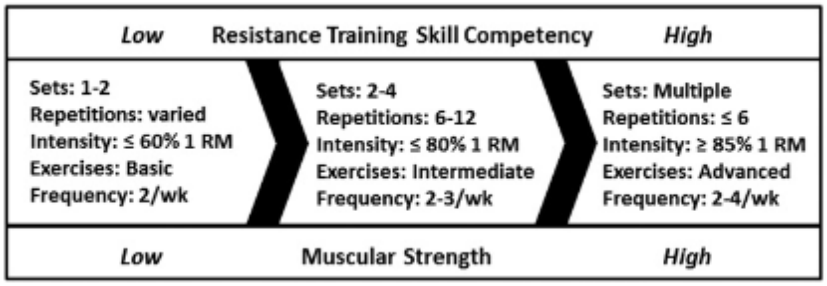



Figure 2. Youth resistance training guidelines with progression based on each participant's resistance training skill competency and muscular strength (Faigenbaum & McFarland, 2016)

Furthermore, strength training should be incorporated to training prior to power training in order to establish an adequate foundation of strength for power training such as jump training (Behm et al. 2017). During prepubescence strength, fundamental movement skills, speed and agility should be focused on and during adolescence sport-specific skills, power, and hypertrophy should be added (Lloyd & Oliver, 2012). Strength training should be developed alongside mobility training, since if the focus is only on the development of mobility with static and passive stretches, it will be negatively reflected in the gymnast's movement control and power output (Osmala et al. 2021).

## 2.3 The importance of strength testing

Testing muscle strength is important for various reasons such as health promotion and injury free participation in the sport as well as to help coaches to monitor the development of their athletes.

Testing strength is an important performance parameter. (Salse-Batán et al. 2022; Osmala et al. 2021; Gasparetto, 2022)

The assessment of physical fitness (PF) is important in gymnastics since it promotes healthy and injury free participation as well as talent identification. Additionally, it also helps coaches and trainers to monitor the development of their athletes. PF consists of speed, strength, endurance, agility, flexibility, balance, and power, which are all requirements in gymnastics. (Salse-Batán et al. 2022)

Strength can be assessed either statically where the muscle contraction is typically isometric or dynamically where the muscle contraction is either concentric or eccentric (Faigenbaum & Micheli, 2017; Reed & Bowen, 2009). Tests allowing only few (<3) repetitions before reaching momentary muscular fatigue are considered to measure strength, and tests where numerous (>12) repetitions are performed before momentary muscular fatigue are considered to measure muscular endurance. In addition, the performance of maximal repetition range also can be used to assess strength. (Faigenbaum & Micheli, 2017)

Laboratory tests are seen as the gold standard when assessing PF, however these tests are expensive and require trained experimenters, which is difficult to execute in the gymnastics context. In comparison, field-based tests are widely recommended since they involve minimal equipment and minimal costs that are easy to administer as well as allow a larger number of gymnasts to be tested in a short period of time. (Salse-Batán et al. 2022)

Clinical tests are an important assessment tool to assess human movement and function. Tests assessing strength such as the standing heel raise test, the functional lower extremity evaluation (FLEE), and the single-leg squat can be reliable for clinical use, especially when performed by an experienced tester (Haitz et al. 2014; Barnett et al. 2015; Räisänen et al. 2016). Clinical tests are easy to use and require minimal costs, but often are only valid when used by a trained professional. Visual rating and assessment done by physiotherapists is a valid tool when assessing young athletes (Whatman, Hume & Hing, 2013).

PF related with health and sport performance, is directly associated with muscle strength improvement (Gasparetto, 2022). In addition, testing sport-like strength is important for athletes and young gymnasts and measuring muscle is considered as an important performance parameter (Osmala et al. 2021; Gasparetto, 2022). There is a need for meaningful, reliable, and sensitive outcome gymnastic-specific fitness tests (Salse-Batán et al. 2022).

In the light of research, testing and observing athletes' performance has positive effects. Testing the strength of athletes, along with other qualities such as balance, proprioception, mobility, can influence the onset of injuries in a preventive way (Thomas & Thomas, 2019). By developing communication between athletes and coaches by actively involving the athletes in training planning, the incidence of injuries can be significantly affected for the better (Kolar et al. 2017).

## 2.4 Strength tests and test-batteries in gymnastics

Based on the results of the literature search, the most studied gymnastic modalities are artistic and rhythmic gymnastics. Even though AGG has similarities with rhythmic gymnastics, journals have very little knowledge regarding the strength assessment or strength training of aesthetic group gymnasts.

Regarding strength assessment, test batteries such as FIG analyses (International Gymnastics Federation), the gymnastics functional measurement tool (GFMT) and the talent opportunity program (TOPS) are commonly used in artistic and rhythmic gymnastics. In addition, dynamometry tests, that measures force, and jump tests such as the vertical jump and long jump tests are the most used to measure strength in gymnastic modalities. However, these methods are mainly used to identify and select sport talents even though they have only proved to be efficient in measuring general physical abilities in gymnastics. (Gasparetto et al. 2022)

In comparison, another review stated several tests to assess gymnasts’ fitness such as the side split test, the handstand test, the vertical jump test, the 20-m run test, the agility test, and the aerobic gymnast anaerobic test. From these recommended tests, only the handstand test tested muscle strength. In terms of the reliability of the data, only four out of sixteen test-retest studies analysed in this review reported validity. (Salse-Batán et al. 2022)

Furthermore, it is important to clarify that the test-retest protocols were not adequate in several studies. Test-retest is the first step when building a test battery, in order to be reliable, strength tests need to also be repeatable. Test-retest reliability refers to the consistency within the measurements when they are repeated, for example when strength tests are being tested with the same subjects more than once under the same circumstances. If correlation of these testing times is high, that is considered evidence for good test-retest reliability (Lindberg et al. 2022; Collins, 2007.) If tests are used before being evaluated reliable or unreliable, it is not clear what is being tested and why. This means that the testing would not be evidence-based. In addition, the test-retest procedures in the research are not defined, therefore reliable test-retest studies would have a clear procedure and guidelines to minimise any possible error.

Finally, a multicenter study of test-retest reliability showed moderate to strong test-retest reliability for all strength and power-related tests in the study such as countermovement jump (CMJ), squat jump (SJ), jump and reach, 20-m sprint, 1-repetition maximum squat, sprint cycling and seated leg press (Lindberg et al. 2022).

Rhythmic gymnastics being closer to AGG than artistic gymnastics, studies highlighting tests for rhythmic gymnasts were focused on more closely. Tests for rhythmic gymnasts included flexibility, balance, mobility, strength, muscular power, cardiorespiratory fitness, and coordination. Tests or test batteries focusing on strength included either strength or power tests such as push-up test, sit-up test, back extension test, and different type of vertical jump tests such as CMJ, SJ, DJ, standing long jump or the FIG test battery. Based on the findings, jump tests and plyometric exercises are the most popular to test and train gymnasts. (Gasparetto et al. 2022; Salse-Batán et al. 2022; Lindberg et al. 2022; Dallas et al. 2021; Deley et al. 2010; Nitzsche et al. 2021)

## 2.5 Strength training in gymnastics

Even though functional training and strength training is recommended to help improve the physical performance of gymnasts, strength training programs used in gymnastics are not clearly defined in journals and terms such as traditionally used general exercises and all-around training are used (Sun, 2023; Xiao, 2023; Zhuang et al. 2023). Journals also highlight lower limb strengthening exercises and jumps such as jumps with two feet and squat jumps. Lower limb exercises have been proven to improve knee strength, which improved technical skills. (Dallas et al. 2021)

Furthermore, isokinetic tests for the knee extensor muscle and vertical jump tests such as the squat jump, counter movement jump and a reactivity test of 6 consecutive jumps have been used in gymnastics (Deley et al. 2010). Isokinetic strength and jump performance of youth rhythmic gymnasts can be improved with plyometric training and should be included to the traditional training to improve the performance of rhythmic gymnasts (Nitzsche et al. 2021).

Because of the lack of definitions in the strength exercises used in the journals, it is clear to see that having a reliable and sport specific testing battery in AGG would be beneficial. Therefore, athletes needing muscle strength could be first identified and then monitored throughout the season.

## 2.6 Summary of the research

The literature search highlighted there is very little knowledge regarding strength training of aesthetic group gymnasts as well as strength assessment. Test-retest studies assessing strength testing in gymnastics were lacking validity. However, the research states that strength training in children has proven to be safe and efficient when properly designed which has beneficial effects for health, strength, and athletic performance.

Measuring muscle strength is an important performance parameter. Testing the strength of athletes, along with other qualities such as balance, proprioception, mobility, can influence the onset of injuries in a preventive way. The strength tests focused heavily on jump and plyometric tests. Therefore, there is a need for meaningful, reliable, and sensitive outcome gymnastic specific fitness tests with a need to test sport like strength tests. This means there is a need for strength tests designed specifically for AGG with the purpose to measure gymnasts sport specific strength and performance. The research proves the need for reliable strength tests in AGG.

The nine strength tests (Appendix 2) evaluated in this research have been developed by the experts of the Finnish Gymnastics Federation to highlight the role of strength needed in the sport. The strength tests have been created for the coaches to use as part of normal training, which will benefit and support the athletes on a daily basis.

# 3 Purpose and objectives

This study evaluates the inter- and intra-rater reliability of the nine strength tests developed by the Finnish Gymnastics Federation.

The research question is the following:

1. Are the strength tests of the Finnish Gymnastics Federation reliable?

The additional research questions are:

1. What is the inter-rater reliability of the Finnish Gymnastics Federation's strength tests?
2. What is the intra-rater reliability of the Finnish Gymnastics Federation's strength tests?

Null hypothesis would be that the strength tests of the Gymnastics Federation are not reliable and repeatable, and the alternative hypothesis would be that the strength tests of the Gymnastics Federation are reliable and thus repeatable.

## 3.1 Purpose and benefits of the research

The purpose of this study is to evaluate the inter- and intra-rater reliability of the strength tests with a test-retest. This evaluation is the very first step in building an evidence-based test battery to monitor athletes in any sport. This paper highlights the role of strength and strength training in gymnastics, aesthetic sports, and in injury prevention, and how it is beneficial from a young age, as well as a tool for coaches to use to evaluate strength in daily situations.

This research has many long-term benefits. Testing athletes' performance is an important part of an athlete's every day to achieve goals but also is in the athletes' best interest to do so. After evaluating the reliability and validity of the strength tests, one of the goals in this study is to highlight the importance of strength characteristics among young gymnasts, especially for injury prevention and performance development. By monitoring and developing performance, we can influence the health of a young athlete and prepare the athlete for the diverse requirements of the sport.

The development of strength and power as a part of the aesthetic disciplines of gymnastics is becoming even more proliferating. It is necessary to critically evaluate the reliability of the strength tests as they have not been done before now. Strength tests are already being used on a daily basis without any analysis of the tests itself. Therefore, it is essential that this research paper critically evaluates the strength tests used in artistic group and rhythmic gymnastics.

## 3.2 Objectives of the study

The need for this research arose from the field. The goal for the Gymnastics Federation was to find out whether the strength tests they developed for aesthetic and rhythmic gymnasts reliably and repeatably measures the performance and strength characteristics.

The target group in this study is a group of under 15-year-old aesthetic group gymnasts, and the intervention is the strength tests developed by the Finnish Gymnastics Federation. The aim of the designed tests is to measure gymnasts' sport-specific strength and performance.

This test-retest evaluates the consistency of the results over time and between testers, in this case on two different testing days and between two independent testers. A test-retest is the most common measure of reliability.

# 4 Material and methods

## 4.1 Study design

The study design in this research is a test-retest evaluating the Gymnastics Federation's strength tests testing inter- and intra-rater reliability.

### 4.1.1 Test-retest

Test-retest reliability refers to the consistency within the measurements when they are repeated, for example when strength tests are being tested with the same subjects more than once under the same circumstances. If correlation of these testing times is high, that is considered evidence for good test-retest reliability (Lindberg et al. 2022; Collins, 2007.) The more testing sessions and the more time between the test sessions there are, the better validity of the data (Lindberg et al. 2022). Still, it has been proved that there was no statistically significant difference in a test-retest reliability at either two days or two weeks apart mark (Marx et al. 2003).

Furthermore, a systematic review stated five test-retest studies with time frames of 24-48h, one week, two weeks, two and a half weeks, and four weeks. Within this review, the one-week timeframe was mostly used (Muñoz-Bermejo et al. 2021). Timeframes for test-retest studies are not clearly defined in journals, for example in one test-retest study athletes performed twice in a three-week period (Tayech et al. 2018). Therefore, journals not stating the conditions of a test-re-test makes it hard to define the optimal test-retest timeframe.

The quality of any research can be downgraded if there are limitations in the study design, the implementation of the study, variability in results, or indirectness of evidence. Therefore, it is important to provide nonbiased, transparent summaries with critical outcomes. (Guyatt et al. 2008)

## 4.2 Methods

The evaluation was done with a single blinded test-retest that evaluated the reliability between two testing days and two testers. The inter- and intra-rater reliability was evaluated with a test-retest in June 2022. For the recruitment of the participants, many different AGG clubs and coaches in Finland were contacted to find volunteers for this study. This test-retest testing was carried out

on two testing days by two physiotherapists. Both the gymnasts and their parents were informed about the study and parental written consent was obtained before participation.

To help formulate the research questions a PICO was used to clearly define the population, the measurement tool, the timings of the test-retest, and to state an outcome measure (Table 1). The population of this study consisted of nine (N=9) under 15-years-old aesthetic group gymnasts. The nine strength tests measuring sport specific strength and performance were the measurement tool. This study evaluated the reliability of the strength tests with a two-day time interval between the testing days. The consistency between the scores obtained from the two testers of the strength tests were measured using statistical measures such as Cronbach's alpha and intraclass correlation coefficient (ICC). All procedures were in accordance with the ethics of the University of the Applied Sciences of Jyväskylä, Finland. The ethical review for the study was approved in February 2022.

Table 1. PICO

| **PICO** |
| --- |
| Population: Under 15-year-old female aesthetic group gymnasts |
| Intervention: Nine strength tests measuring sport specific strength and performance as a measurement tool |
| Comparison: Two-day time interval between the testing days |
| Outcome: The consistency between the scores obtained from the two testers of the strength tests using statistical measures such as Cronbach's alpha and intraclass correlation coefficient (ICC) |

The testing was done twice within the same timeframe and in the same location under the same circumstances on both days as a part of the gymnasts' normal training. This allowed the reliability and repeatability of the tests to be assessed by the two independent testers when having identical testing situations on both testing days. The test-retest was executed two days apart on a Monday and a Wednesday, between 9am – 3pm.

The participants were selected randomly using the inclusion and exclusion criteria seen in Table 2. Forms were sent to the club and parents two weeks prior to the testing part of this study. The forms included a consent form and a preliminary information form as well as an in-depth insight to the study which provided the gymnast's and their guardians all information needed regarding this study and participation to it. Selection of the participants and the randomisation of the subjects was done manually at the start of the testing. If the collected forms did not meet the inclusion criteria, they were excluded. Nineteen (N=19) forms were randomly collected and screened through, and nine (N=9) gymnasts were included into the study.

Table 2. Inclusion and exclusion criteria

| **Inclusion criteria** | **Exclusion criteria** |
|---|---|
| Consent | No consent |
| Under 15-years-old female | Over 15-years-old female |
| Aesthetic group gymnast | Not an aesthetic group gymnast |
| No injury or illness during the previous 6 months | Injury or illness during the previous 6 months |
| Completed and signed consent and information forms | Missing signatures or incomplete forms |

Before the start of the testing, the testing location in the training location was clearly defined with the equipment needed. All gymnasts were met individually and given the same warmup prior to the testing. After approximately 10 minutes of warmup, gymnasts were asked to state their exhaustion level using a Borg CR10 RPE scale to assess their level of readiness (Martínez & Grande, 2021; Fairman et al. 2018). Appendix 3 states the warmup and the RPE scale used in the testing.

Testers were blinded and did not know each other's scores throughout the testing days. Scores were kept hidden until the analysis of the data was completed. Bias was limited by testing the gymnasts in the same order, in the same location, and within the same timeframe. All gymnasts received the same treatment, guidelines, and times to rehearse the test movements prior to testing. No scores were revealed to the gymnasts, to the coaches, or between the testers.

## 4.3 The strength tests and the testing protocol

The strength tests consist of nine tests that measure the sport-specific strength and performance of gymnasts. The tests, the guidelines, and the scoring system have been developed by the experts of the Finnish Gymnastics Federation.

The test battery includes nine different strength tests. The strength tests are evaluated by using a numeric score of zero, two, four with zero (0) being the lowest and four (4) being the highest score. Each test movement has a specific criterion for each score. For example, test three measuring hip flexion and core strength requires ten leg lifts for full points (4/4) and five leg lifts for the second highest score (2/4). Zero points will be given if the subject is either not able to get to the test position or does less than five leg lifts in total. The testing protocol and the scoring varies between the nine strength tests. Based on the test, the highest score requires either multiple repetitions, a hold with required control and alignment, one repetition or an execution of a movement. Based on the test protocol and execution, some of the strength test highlight control, some sport like performance and some focus purely on strength. The original strength tests are in Finnish and can be found in Appendix 2. However, an English summary of the strength tests is shown in Table 3.

Table 3. Summary of the strength tests

| | Test names | Test equipment | Requires assistance | Full points (4/4) requirement | Type of performance |
|---|---|---|---|---|---|
| 1. | Glute muscles | Tape | Yes (fixation of the pelvis) | Hip extension with straight leg in prone position high-lighting body control | 5 second hold |
| 2. | Adductor muscles | Gymnastic bench, tape, wooden stick | Yes (stick used to stop pelvis rota-tion) | Copenhagen -type of movement: Lower leg lift in side lying with pelvic control | 5 second hold |
| 3. | Hip flexors | Wall bars | No | 10 leg lifts above head | Many repetitions |
| 4. | Shoulders in hand-stand push-up | - | No | Controlled push-up in hand-stand position | One repetition |
| 5. | Lateral hip strength | - | Yes (fixation of the pelvis) | Bended knee hold in prone position with good control while maintaining position | 5 second hold |
| 6. | Rotation of the core | - | No | Controlled leg lift using both legs highlighting the rotation strength of the core | One repetition |
| 7. | Hamstrings | - | Yes (holding the legs down) | Nordic hamstring -type of movement | One repetition |
| 8. | One leg heel raises | Metronome, wooden stick | Yes (one hand in front of the knee) | >30 one leg heel raises with good control | Many repetitions |
| 9. | Back extension | Two wooden sticks | Yes (holding the legs down) | Maximal back extension | One repetition |

This study conducted a single-blinded test-retest for a group of participants where the subjects were selected based on inclusion and exclusion criteria and put in a randomised testing order. Both the subjects and testers were blinded by the grading system. The strength tests used were not taught beforehand, nor was the subjects or their guardians given any educational material re-garding the tests itself. The positive in the athlete's not knowing the tests beforehand, could make the testing more reliable, whereas the negative of the athlete's not knowing the tests, could be a possible reduction in test performance.

To avoid bias the testing was executed by two independent testers and the testing was completed single blinded. The two testers did not see each other's scores and the gymnasts did not know their results on either of the two testing days. In addition, the testing was executed in the same location within the same timeframe on both testing times. The testers were both physiotherapists with backgrounds in sport physiotherapy.

19 gymnasts were recruited to participate in the study, but only 9 were included based on the inclusion and exclusion criteria as shown in Table 2. The testing order was randomly formed when collecting the forms. All the gymnasts were explained the inclusion and exclusion criteria in the beginning of the testing. Each participant was called individually to the testing location.

In the beginning of the testing a pre-designed warmup was instructed to each gymnast. The warmup included movements such as jogging, squats, lunges, inch worms, bear walks, core, and back exercises (Appendix 3). After the warmup the Borg-Scale was used to assess the gymnasts' level of readiness to start the testing and an RPE score of 7 out of 10 was required to start the testing. In the beginning of the test, each test movement was explained and instructed verbally, and the gymnasts were able to do a practice test movement 1-2 times before the test itself. The testing was completed on the side of a large gymnastic training hall as a part of their normal training between 9am and 3pm.

## 4.4 Collection and description of data

### 4.4.1 Collection of the data

The research data collected in the study on the gymnasts participating includes the following information: age, gender, height, weight, training history, injury history, current skill, and training level. The athlete demographics are presented in the results (Table 5). The research data was first collected on paper in written form, after which the research data was digitised, and the research subjects received a random personal identity code. The digital research data does not contain identifiable metadata. The list of personal data is stored in a locked cabinet by the researcher and Jyväskylä University of Applied Sciences. Digital data is stored in encrypted excel (.xlsx), word (.docx) and SPSS documents (.sav). All research data will be preserved throughout the project. All data is collected during the project and is less than 1GB.

On the testing day out of nineteen (19) gymnasts nine (9) met the criteria – two (3) had missing signatures/forms and seven (7) gymnasts were injured during the past 6 months. The testing of the study took place in the gymnasts own training facilities, making it as easy as possible for the gymnasts to participate in the study as a part of their own training. Nine (9) gymnasts qualified to participate in this study and all 9 gymnasts participated on both testing days.

### 4.4.2 Description of data

A numerical scoring system was used to score the strength tests. Each of the nine strength tests were evaluated with 0, 2, 4 with 0 being the lowest score and 4 the highest score. Data included scores from nine gymnasts on nine different strength tests and was scored by two different testers on two different testing days. The combined data included the scores from left and right for unilateral exercises. In the strength tests 1, 2, 5, 6 and 8, evaluating unilateral strength, the side with the higher scores was included into the combined data to measure and see the athletes' strengths rather than the weaknesses. Using the stronger side in the unilateral strength tests shows the athletes' current levels of strength, which needs to be established to understand how strong the athletes currently are. In addition, using the maximum scores gave definitive results compared to the averages of both side in the unilateral strength tests. The data collected in writing is encoded in the analysis phase in encrypted Excel (.xlsx) and SPSS documents (.sav).

Results of the individual strength tests were classed as categorical data and results of the athlete's total scores as continuous data. Statistical analyses of the data were carried out using the SPSS program (IBM SPSS Statistic) and Microsoft Excel. For the statistical analyses Cronbach's alpha was used for the categorical data and the intraclass correlation coefficients (ICC), the independent samples t-test and Bland-Altman plot analysis were used for the continuous data.

Cronbach's alpha being a measure of internal consistency, it was used to measure the internal consistency of the individual strength tests, to find out the extent to which all the items in a test measure the same concept or construct (Tavakol & Dennick, 2011). Whereas the ICC was used to investigate the reliability of this test-retest with the athlete's total scores. The ICC was based on a two-way mixed-effects model and an absolute agreement (Koo & Li, 30 2016). The independent samples t-test was used to compare the tester's mean values of the athletes total scores on both test-retest days, and the Bland-Altman plot evaluated a bias between the mean differences of the

testers and the testing days (Xu et al. 2017; Giavarina, 2015). A Bland-Altman plot displays a relationship between two paired variables using the same scale. It is a scatter plot where the differences between two measurements are plotted against their averages (Giavarina, 2015). The limit of statistical significance was defined as the p-value (< 0,05).

## 4.5 Definitions of reliability

### 4.5.1 Reliability

Reliability is defined as the accuracy of an instrument and relates to the consistency of a measure. Reliability is a second measure of quality used in a quantitative study, and what this test-retest focuses on. This means the extent to which a research instrument in this case, a strength test, has consistently the same results if used in the same situation or repeated occasion. A tester executing the strength test should have approximately the same results each time the tests are completed. Homogeneity, stability, and equivalence are the three attributes of reliability. (Heale & Twycross, 2015)

The Cronbach's α is the most used test to determine internal consistency, homogeneity, of an instrument. The Cronbach's α result is a number between 0 and 1, acceptable reliability score being 0.7 and higher. Stability can be tested with a test-retest and equivalence with inter- and intra-rater reliability. To get good reliability, consistency of results across time (test-retest), across different observes (inter-rater) and across paths of the tests itself (intra-rater) should be high. (Heale & Twycross, 2015)

### 4.5.2 Inter- and intra-rater reliability

Inter-rater reliability refers to two or more individuals who observe or measure the same situation. Inter-rater reliability is seen as perfect when two or more individuals agree on all items after observing individually. Intra-rater reliability refers to the consistency of measurement of an individual. Inter- and intra-rater reliability are aspects of test validity and if seen high the intraclass correlation coefficient is high. (Fink, 2010; Heale & Twycross, 2015)

### 4.5.3 Intraclass correlation coefficients and Cronbach's alpha

Intraclass correlation coefficients (ICC) is a common and widely used reliability index in test-retest, inter-rater and intra-rater reliability analyses (Koo & Li, 2016). ICC is a value between 0 and 1, where 0.5 and below indicates poor reliability, between 0.5 and 0.75 indicates moderate reliability, between 0.75 and 0.9 indicates good reliability, above 0.9 indicates excellent reliability based on the 95% confident interval of the ICC estimate (Bobak et al. 2018; Koo & Li, 2016). In this test-retest the ICC is used to evaluate the athlete's total scores between the two testers and the two testing days.

Cronbach's alpha is a measure of internal consistency, which shows how closely related a set of items are as a group. Also, it is a measure of scale reliability. Table 4 states an acceptable value is anything between 0.70 to 0.95 with 0.9 being excellent (Tavakol & Dennick, 2011; Bobak et al. 2018; Koo & Li, 2016). In this test-retest the Cronbach's Alpha is used to measure internal consistency of the nine strength tests.

Table 4. Internal consistency (Bobak et al. 2018; Koo & Li, 2016)

| **Cronbach's alpha** | **Internal consistency** |
|---|---|
| $\alpha \geq 0.9$ | Excellent |
| $0.9 > \alpha \geq 0.8$ | Good |
| $0.8 > \alpha \geq 0.7$ | Acceptable |
| $0.7 > \alpha \geq 0.6$ | Questionable |
| $0.6 > \alpha \geq 0.5$ | Poor |
| $0.5 > \alpha$ | Unacceptable |

## 4.6 Ethical assessment

The ethical principles of this study were met and assessed prior to the start this test-retest study. Ethical assessment was required due to the population consisting of a group of underage aesthetic group gymnasts. The degree regulations of Jyväskylä University of Applied Sciences are rules that are approved by the academic board. The ethical approval of this study was assessed by the ethics

committee of Jyväskylä University of Applied Sciences in August 2021 and approved on the 1st of February 2022.

### 4.6.1 Ethical principles of this study

This study did not interfere with the physical integrity of the subjects, but merely evaluated the sport-specific test movements that the participants performed in the testing situations. The researcher respects the dignity of the research participants and their right to self-determination both during and after the research (1999/731, sections 6–23). The research was carried out in a way that does not cause significant harm, damage or risk to the participants, communities or other parties involved. (TENK, 2021)

Participants and their parents were informed about the study and its practices in written form. Written consent to participate in the study was requested from the participants or their parents, depending on the age of the participant. In any case, parents and guardians were informed about the content and course of the study. If the participant had already reached the age of 15, they were allowed to decide on their own participation. The participants recruited for the study had full decision-making power to participate and not participate in the study. Voluntary consent to participate in the study was requested from the participants and guardians in writing. (TENK, 2021)

Participants had the right to withdraw from or discontinue the study at any time. Research subjects had the right to receive information about the content of the research, the processing methods of personal data and how the data used in the research was aggregated. Participants received understandable, truthful, and informative information about the research and its course. It was the researcher's responsibility to inform about the effects and potential benefits of the research. The research subjects had the right to know about the implementation of the research, whether the researcher is present or not. Minors were explained in an understandable way about participation in the study and their role in the research. The guardians of participants under the age of 15 were asked for their consent to participate in the study. Children over the age of 15 can decide upon their own participation, but guardians must still be informed. In all cases, written consent from the participants themselves was required for their participation. (TENK, 2021)

The autonomy of the participants was respected in all situations. The researcher had the right to terminate the participation of a minor participant in the research if it was no longer in the participant's best interest or the participant had stated that they wanted to discontinue their participation. (TENK, 2021)

# 5 Results

The results have been formed from two different types of data: categorical data and continuous data. Categorical data refers to the results of the individual strength tests and continuous data refers to the total scores of the athletes. Cronbach's alpha was used for the categorical data and ICC-values, the independent samples t-test and Bland-Altman plot analysis were used for the continuous data.

## 5.1 Athlete demographics

Table 5. Athlete demographics

| Characteristics | Mean | SD |
|---|---|---|
| Age (years) | 12.33 | 0.50 |
| Height (cm) | 157.48 | 6.15 |
| Weight (kg) | 43 | 5.07 |

The mean and the standard deviation (SD) of the athlete demographics showing their age, height, and weight can be seen in Table 5. In addition, the skill level of all athletes was national level with the experience varying between 2-9 years. Training hours per week consisted of 15-20 hours with only one athlete stating 10-15 hours.

## 5.2 Intra-rater reliability

Table 6. Intra-rater reliability

| **TESTS** | **DAY 1** | **DAY 2** |
|---|---|---|
| | Cronbach's Alpha (α) | Cronbach's Alpha (α) |
| 1 | **0.591** | 0.782 |
| 2 | **0.591** | 0.896 |
| 3 | 0.966 | 1.00 |
| 4 | 0.857 | 1.00 |
| 5 | 0.978 | 0.966 |
| 6 | 0.976 | 0.896 |
| 7 | **0.643** | **0.640** |
| 8 | 0.966 | 0.990 |
| 9 | 0.923 | 0.923 |

Table 6 shows the Cronbach's alpha internal consistency score for all nine strength tests on both testing days, day one and day two. Tests one and two on testing day one showed the lowest score of 0.591, which is seen as a poor consistency score. Internal consistency score is higher on the second testing day where only one test, test seven, is showing a questionable score of 0.640. This means the athletes tested more closely together as a group on the second testing day leading to higher consistency scores.

## 5.3 Inter-rater reliability

Table 7. Inter-rater reliability

| TESTS | TESTER 1 | TESTER 2 |
|---|---|---|
| | Cronbach's Alpha (α) | Cronbach's Alpha (α) |
| 1 | 0.960 | **0.673** |
| 2 | **0.574** | 0.822 |
| 3 | 1.0 | 0.966 |
| 4 | 0.800 | 0.800 |
| 5 | 0.857 | 0.887 |
| 6 | 0.942 | 0.736 |
| 7 | **0.357** | 0.818 |
| 8 | 0.885 | 0.888 |
| 9 | 1.0 | 1.0 |

Table 7 shows the Cronbach's alpha internal consistency score for all nine strength tests between the two testers. Tests two and seven for tester one showed the lowest scores of 0.574 and 0.357 which are seen as poor and unacceptable consistency scores. Tester two's internal consistency score is higher because only test one is showing a questionable score of 0.673. This means tester two had a higher internal consistency on all tests for all athletes on both testing days.

## 5.4 Scores between testing days for all athletes

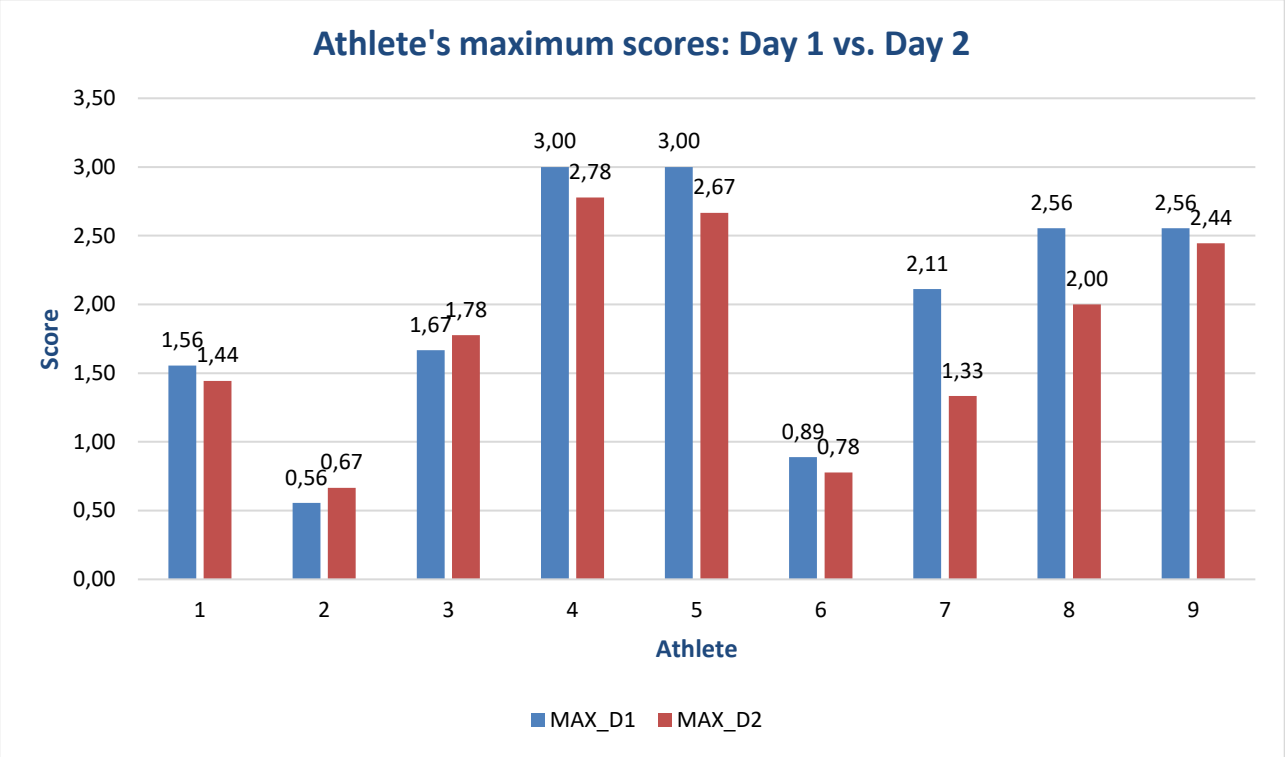



Figure 3. Maximum scores between testing days

Figure 3 shows the average of the maximum score given by both testers on day one and day two for all nine athletes. Four is the maximum score that can be given in each of the nine strength tests. This chart shows a slight difference between the two testing days, showing there is a slight decline in the overall score on the second testing day. This figure shows close agreement between the maximum scores on both testing days.

## 5.5 Athlete's total scores

Table 8. Athlete's total scores on both testing days

| **Athlete's Total Scores** | | | | |
|---|---|---|---|---|
| Test-retest Subjects | Testing day 1 | | Testing day 2 | |
| | Tester 1 | Tester 2 | Tester 1 | Tester 2 |
| 1 | 12 | 16 | 14 | 14 |
| 2 | 4 | 6 | 8 | 8 |
| 3 | 16 | 16 | 16 | 16 |
| 4 | 24 | 30 | 24 | 28 |
| 5 | 30 | 28 | 24 | 24 |
| 6 | 10 | 10 | 8 | 8 |
| 7 | 16 | 20 | 14 | 12 |
| 8 | 20 | 24 | 20 | 16 |
| 9 | 18 | 24 | 24 | 24 |
| The total average of the maximum 36 points | 16,5/36 | 18,75/36 | 16/36 | 15,75/36 |

The athlete's total scores in Table 8 show the combined scores from all nine strength tests for each athlete. In the unilateral strength tests, the side with the higher score were used as a maximum score to highlight the athletes' stronger side instead of the athletes' weaker side. The total scores seen in this table are from a total of 36 points. There can be seen that the total scores of athletes are slightly higher on the first day compared to the second day. In addition, there can be seen an average score difference of 0.5 points between tester one and tester two on testing day one, and 3 points difference on testing day two.

## 5.6 Bland-Altman plot for all tests total

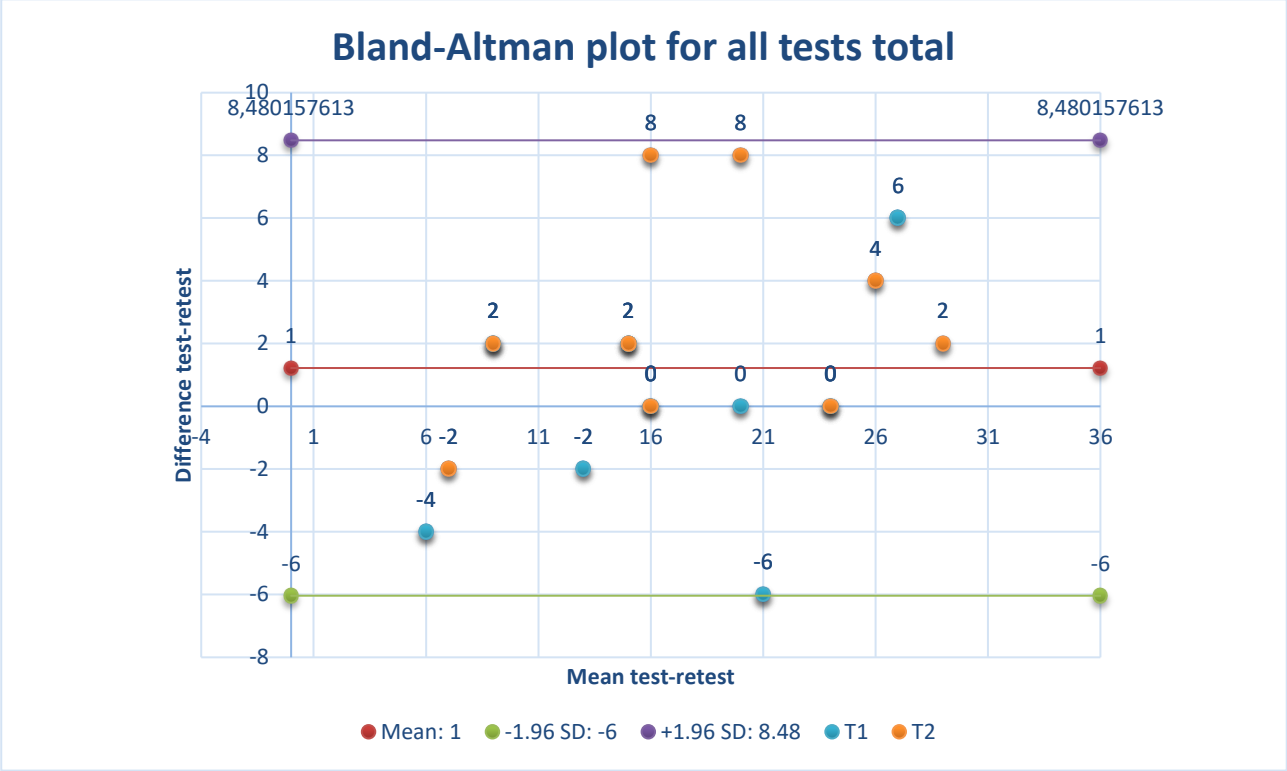



Figure 4. The Bland-Altman plot for all tests total

This Bland-Altman plot is used to illustrate absolute reliability and the consistency of gymnasts' total scores using their stronger side. This graph shows all athlete's total scores between the two testing days of the test-retest as well as between the two testers. This graph, Figure 4, displays the relationship between two paired variables using the same scale. In this scatter plot, the differences between two measurements and the two testing days are plotted against their averages. Here both testers, tester one and tester two, are measuring the same parameters, the athlete's total scores. There can be seen an agreement between the measurements of two testers which means the athlete's total scores correlate with each other's showing consistent variability. The mean difference is small being close to zero and most of the data points are scattered closely to the mean without any specific trend.

## 5.7 Correlation between the athlete's maximum total scores

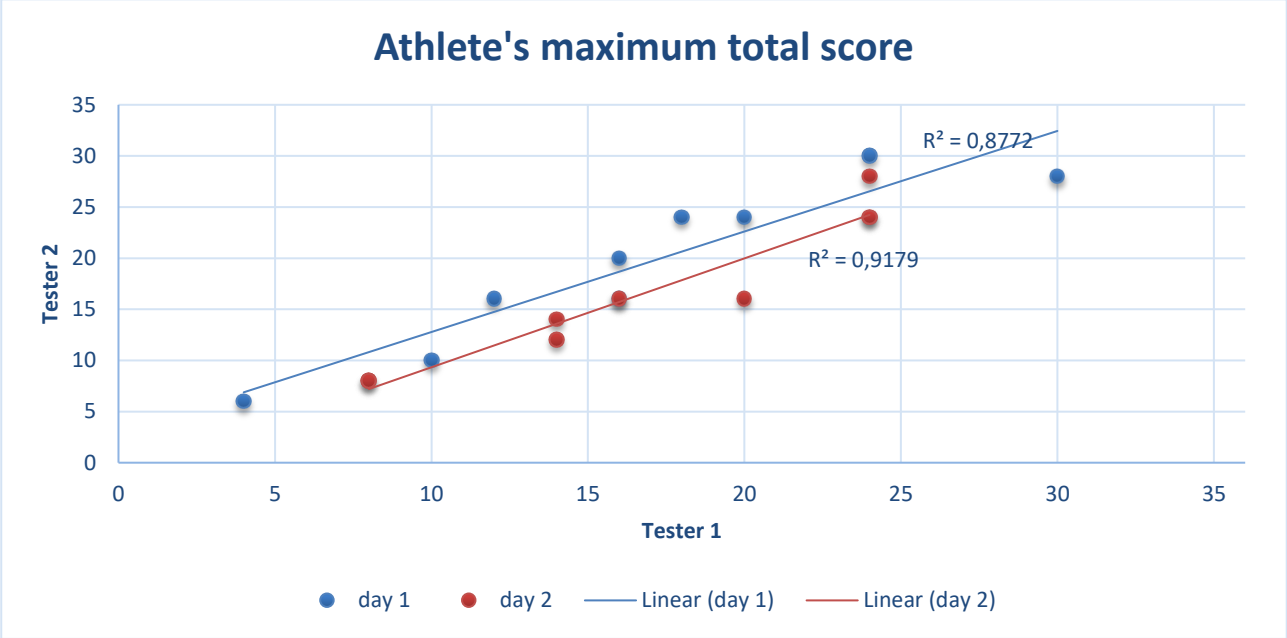



Figure 5. Correlation between the athlete's total scores

The Figure 5 shows the athletes maximum total scores from all nine strength tests. The total scores are shown by both testers. The x-axis shows tester one's total scores and the y-axis shows tester two's total scores from all nine athletes. Correlation of the athlete's total scores can be identified as high between the two testers and the two testing days. The athlete's total scores correlate with each other. The actual overall strength scores identify those athletes with lower and higher overall strength. R-squared values for both testing days can be seen as high. $R^2$ =0.8772 on testing day one and $R^2$ = 0.9179 on testing day two. $R^2$ value over 0.9 is seen as high and further verifies the correlation between the athlete's total scores.

## 5.8 The independent samples t-tests

### 5.8.1 The independent samples t-test for tester one

A two-sample t-test was performed to compare the athlete's results of tester one between testing day one and testing day two (Figure 6 and Figure 7). There was not a significant difference in the athlete's results of tester one between testing day one (M = 16.66, SD = 7.68) and testing day two (M = 16.88, SD = 6.49); t(16) = -.066, p = .948. Since the p-value of the test (.948) is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the results of tester one is different between testing day one and testing day two.

**Group Statistics**

| | DAY | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| TESTER 1 | 1,00 | 9 | 16,6667 | 7,68115 | 2,56038 |
| | 2,00 | 9 | 16,8889 | 6,48931 | 2,16310 |

Figure 6. T-test group statistics tester one

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Significance | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | One-Sided p | Two-Sided p | Mean Difference | Std. Error Difference | Lower | Upper |
| TESTER_1 | Equal variances assumed | ,011 | ,918 | -,066 | 16 | ,474 | ,948 | -,22222 | 3,35180 | -7,32772 | 6,88328 |
| | Equal variances not assumed | | | -,066 | 15,566 | ,474 | ,948 | -,22222 | 3,35180 | -7,34387 | 6,89942 |

Figure 7. Independent samples t-test tester one

### 5.8.2 The independent samples t-test for tester two

A two-sample t-test was performed to compare the athlete results of tester two between testing day one and testing day two (Figure 8 and Figure 9). There was not a significant difference in the athlete's results of tester two between testing day one (M = 19.33, SD = 8.06) and testing day two (M = 16.66, SD = 7.21); t(16) = .740, p = .470. Since the p-value of the test (.470) is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the results of tester two is different between testing day one and testing day two.

**Group Statistics**

| | DAY | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| TESTER_2 | 1,00 | 9 | 19,3333 | 8,06226 | 2,68742 |
| | 2,00 | 9 | 16,6667 | 7,21110 | 2,40370 |

Figure 8. T-test group statistics tester two

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Significance | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | One-Sided p | Two-Sided p | Mean Difference | Std. Error Difference | Lower | Upper |
| TESTER_2 | Equal variances assumed | ,156 | ,698 | ,740 | 16 | ,235 | ,470 | 2,66667 | 3,60555 | -4,97676 | 10,31009 |
| | Equal variances not assumed | | | ,740 | 15,805 | ,235 | ,470 | 2,66667 | 3,60555 | -4,98444 | 10,31777 |

Figure 9. Independent samples t-test tester two

### 5.8.3 The independent samples t-test for testing day one

A two-sample t-test was performed to compare the athlete's results on testing day one between tester one and tester two (Figure 10 and Figure 11). There was not a significant difference in the athlete's results on testing day one between tester one (M = 16.66, SD = 7.68) and tester two (M = 19.33, SD = 8.06); t(16) = -.718, p = .483. Since the p-value of the test (.483) is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the results on testing day one is different between tester one and tester two.

**Group Statistics**

| | TESTER | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| DAY_1 | 1,00 | 9 | 16,6667 | 7,68115 | 2,56038 |
| | 2,00 | 9 | 19,3333 | 8,06226 | 2,68742 |

Figure 10. T-test group statistics testing day one

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Significance | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | One-Sided p | Two-Sided p | Mean Difference | Std. Error Difference | Lower | Upper |
| DAY_1 | Equal variances assumed | ,175 | ,681 | -,718 | 16 | ,241 | ,483 | -2,66667 | 3,71184 | -10,53542 | 5,20209 |
| | Equal variances not assumed | | | -,718 | 15,963 | ,241 | ,483 | -2,66667 | 3,71184 | -10,53692 | 5,20359 |

Figure 11. Independent samples t-test testing day one

### 5.8.4 The independent samples t-test for testing day two

A two-sample t-test was performed to compare the athlete's results on testing day two between tester one and tester two (Figure 12 and Figure 13). There was not a significant difference in the athlete's results on testing day two between tester one (M = 16.88, SD = 6.49) and tester two (M = 16.66, SD = 7.21); t(16) = .069, p = .946. Since the p-value of the test (.946) is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the results on testing day two is different between tester one and tester two.

**Group Statistics**

| | TESTER | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| DAY_2 | 1,00 | 9 | 16,8889 | 6,48931 | 2,16310 |
| | 2,00 | 9 | 16,6667 | 7,21110 | 2,40370 |

Figure 12. T-test group statistics testing day two

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Significance | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | One-Sided p | Two-Sided p | Mean Difference | Std. Error Difference | Lower | Upper |
| DAY_2 | Equal variances assumed | ,046 | ,833 | ,069 | 16 | ,473 | ,946 | ,22222 | 3,23370 | -6,63291 | 7,07735 |
| | Equal variances not assumed | | | ,069 | 15,825 | ,473 | ,946 | ,22222 | 3,23370 | -6,63906 | 7,08351 |

Figure 13. Independent samples t-test testing day two

## 5.9 Intraclass correlation coefficients (ICC) for athlete's total scores

Table 9. The ICC-values for the athlete's total scores

| Variable | ICC | 95% Confidence Interval Lower Bound | 95% Confidence Interval Upper Bound |
|---|---|---|---|
| Tester 1 | 0.889 | 0.579 | 0.974 |
| Tester 2 | 0.855 | 0.387 | 0.967 |
| Testing day 1 | 0.891 | 0.364 | 0.977 |
| Testing day 2 | 0.957 | 0.825 | 0.990 |

The intraclass correlation coefficients for athlete's total scores can be seen in Appendix 4 and a summary in Table 9. ICC-values for the athlete's total scores varied between testers 0.880-0.855 and between testing days 0.891-0.957 meaning good to excellent reliability based on the 95% confident interval of the ICC estimate.

## 5.10 Summary of the results

To summarise the results, tester one was shown to be more reliable between the two testers and the results on the second testing day were more reliable compared to the first testing day. In addition, there was a slight decline in the overall score on the second testing day meaning the athletes performed with higher scores on the first testing day. When analysing the athletes total scores, there can be seen an agreement between the measurements of the two testers in Figure 4 and Figure 5, showing the athlete's total scores correlating with each other's. Since the p-values of the independent samples t-tests, measuring the two testers and the two testing days, were not less than 0.05, we fail to reject the null hypothesis.

The overall strength tests scores can be used to identify those athletes with lower and higher overall strength. ICC-values for the athlete's total scores varied between testers 0.880-0.855 and between testing days 0.891-0.957 meaning good to excellent reliability based on the 95% confident interval of the ICC estimate.

# 6 Practical applications of this research

Based on this test-retest, the strength test battery can be used as an evidence-based tool for coaches to individualise gymnasts training even further. Due to the multiple demands of the sport, it is necessary that the training is safe and individualised to prevent the multiple injuries occurring in the sport. Figure 14 shows the reasons behind the individualisation tool.

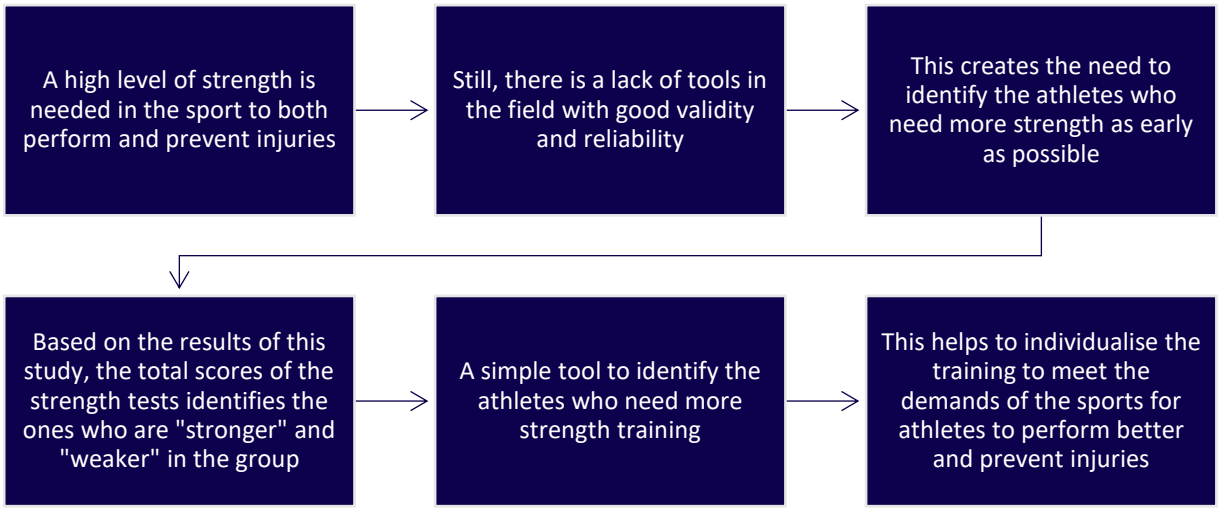



Figure 14. Strength test battery as an evidence-based tool for coaches

Depending on the athlete's overall score from the Finnish Gymnastic Federation's strength tests, athletes can be identified as "stronger" or "weaker". Table 10 shows the scoring system of the athletes. The tool seen in Figure 15 has been developed for the coaches to use in the everyday training. The athletes would be identified based on their overall score. This tool is an additional tool to use on top of normal practices.

The athlete identification tool has been divided into three sections: green, yellow and red. These colours represent percentage of the total score as seen in Table 10. Over 27 points would leave the athlete on green, which classes the athletes as stronger, therefore is ideal for the athlete. Points ranging from 18 to 26 leaves the athletes on yellow, needing more strength training. Points under 17 leaves the athletes on red as "weak" and "immobile". Low levels of strength are dangerous for the athlete and often leads them into difficulties to perform in the sport.

Table 10. The scoring for the athlete identification tool

| % of the total score of the athlete | |
|---|---|
| 75-100% | 27-36 points |
| 50-74% | 18-26 points |
| 0-49% | 0-17 points |

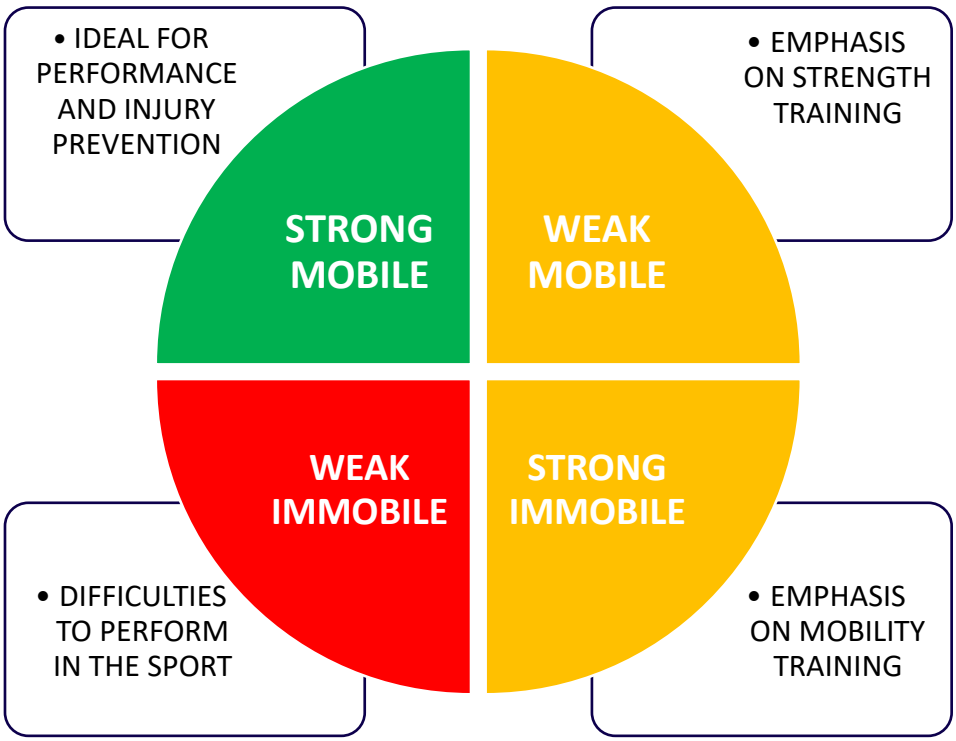



Figure 15. Athlete identification tool

To further analyse the sport-specificity of the strength tests, the following graphic was made (Figure 16). The sports specificity is based on the literature research as well as the testing performed in this study. Tests assessing leg strength and power are beneficial for the performance and gymnastic skills as well as a tool for injury prevention in AGG.

Strength tests three, eight, and nine are proven to be reliable and are evidence-based strength tests, therefore should be kept in the test battery in the future. These three strength tests measure important areas for an aesthetic group gymnast such as leg, hip, and back strength. In addition, tests assessing leg strength and power should be added to the test battery in the future.

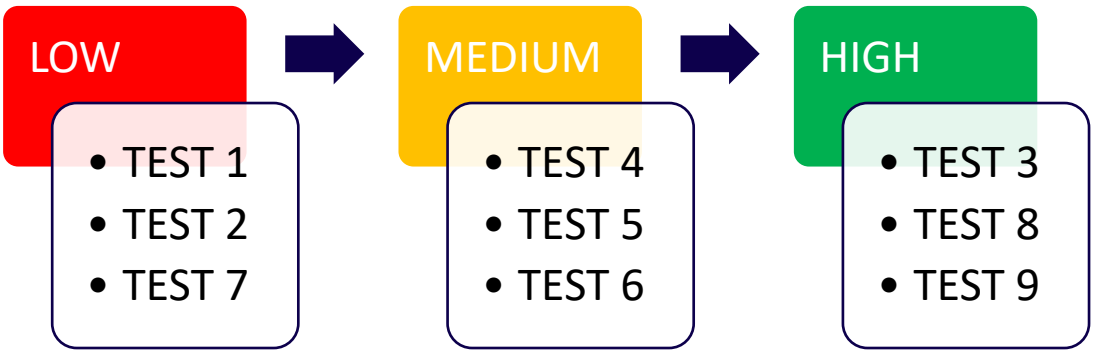



Figure 16. Test sport-specificity

# 7 Discussion

The purpose of this study was to evaluate the inter- and intra-rater reliability of the Finnish Gymnastic Federation's strength tests with a test-retest. It has been established that there is a definite need for reliable strength tests in the field of AGG and rhythmic gymnastics (Salse-Batán et al. 2022; Gasparetto, 2022). This is the first time such a project has been attempted and an important first step in creating a final testing protocol for AGG and strength.

The nine strength tests evaluated in this research have been developed by the experts of the Finnish Gymnastics Federation with the aim to highlight the role of strength needed in AGG and rhythmic gymnastics. The tests are designed to measure gymnasts' sport-specific strength and performance and have been created for the coaches to use as part of normal training. Aesthetic group gymnasts need muscle strength to perform and compete at a high level. A lack of muscle strength at a high level will come at a risk for the athlete. Having strength training and strength testing protocols in place, alongside other physical qualities, will benefit the athletes on a daily basis by preventing injuries and improving performance. Competing gymnasts have a high risk of injury, which became evident early on as a part of the recruitment process, when 7 out of 19 gymnasts (37%) were excluded due to an injury in the previous six months. 7 gymnasts being excluded due to an injury is a high number for both this age group and this sport. (Campbell et al. 2019; Thomas & Thomas, 2019; Alwasif, 2019; Sawczyn et al. 2016; Dallas et al. 2021; Deley et al. 2010; Gasparetto, 2022; Salse-Batán et al. 2022)

Based on the results, the individual strength tests varied in terms of their reliability, but as a strength test battery they can be used to individualise the everyday training of the gymnasts by identifying the "stronger" and "weaker" athletes. This finding is a much needed addition to the athletes everyday training. This research is an important and essential first step when developing strength testing and strength training culture for the sport of AGG.

Furthermore, this test-retest showed clear differences between both the testers and the testing days. Tester one's results were more significant than tester two's (Table 7), and the second testing day had higher internal consistency even though the athlete's overall scores were lower. This means the second testing day had a higher reliability compared to the first testing day (Table 6).

Tester one being more reliable as a tester, but still having slightly lower internal consistency, suggests that there are inconsistencies in the test movements and guidelines of the tests which makes the testing and scoring more vulnerable for error. Testing day two, being more reliable in terms of the results, still had questionable internal consistency in the tests, suggesting that the repeatability of the strength tests is not high enough.

Differences between the testing days can be a result of the athletes or the testers learning from the first testing day. Athletes could have also been practicing the movements between the testing days. Practicing the movements as well as the overall testing protocol could indicate that practicing the tests could be used to improve the consistency of the testing, which would naturally happen as the athletes repeat the tests during the season. Yet, it is important to acknowledge that for a reliable strength test needs to be able to be repeated by anyone at any time (Heale & Twycross, 2015; Lindberg et al. 2022; Collins, 2007).

The variation in the test movements and their guidelines made some of the tests simpler to execute. Especially test number four measuring upper body strength with a handstand turned out to be challenging for the athletes to perform and the testers to score. Majority of the gymnasts struggled coming down from a handstand into a headstand in a controlled manor. Due to the lack of upper body strength and control, a pillow was needed to place under their heads and physical help was required with some due to safety concerns. Assistance was given by slightly holding their legs during the eccentric muscle work to help a safe landing. Any physical help given automatically gave zero points. A handstand push up, starting from a handstand position back against the wall, turned out to be too difficult for the athletes to perform safely. Testing the gymnasts upper body strength might not be needed due to the requirements of the sport but can still be beneficial for the athletes when assessed correctly (Salse-Batán et al. 2022; Osmala et al. 2021; Gasparetto, 2022).

In addition, assessing the execution of the strength tests turned out to be challenging when a tester needed to be hands on while assessing the performance. For example, in test number eight measuring calf strength, the test movement turned out to be challenging to evaluate due to the testers positioning. During the test, a tester counts the amount of calf raises performed with a metronome, monitors the heel height on each heel raise, but should also keep one hand placed on the subject’s knee to help keep it straight. Measuring the heel height is important as the measured

heel height needs to be kept on every repetition, otherwise the test is over. Yet, the fixation of the knee by the therapist is not required in a clinical setting when assessing an athlete, therefore makes it an unnecessary addition to this test battery (Hébert-Losier et al. 2017). The straightness of the knee during the heel raises could instead be used as one of the requirements for the test.

Both testers in this test-retest are therapists with years of clinical experience in physiotherapy. The only difference is the type of sports background with tester one being the researcher and tester two only assisting with the testing. Tester one's background is in aesthetic sports and tester twos in team sports. As a researcher tester one is more familiar with these strength tests, which could lead into better understanding and therefore more reliable scoring over the two testing days. Being more familiar with the tests, the instructions, and the scoring system, could then lead into more reliable results. This could suggest that the strength tests are only reliable if you know the tests and the scoring system of them or the tester has a similar sports background.

When looking at the intra-rater reliability of the strength tests, tests one and two had poor internal consistency and test seven had questionable internal consistency on the first testing day. The second testing day also had inconsistencies in the results, but only test seven had questionable internal consistency leaving the test seven with the lowest overall reliability. Differences in results on both testing days show that the intra-rater reliability is not high enough. Tests one, two, and seven, also required physical assistance from the tester, which may have negatively affected the reliability of the tests.

A high inter- and intra-rater reliability means that the strength tests can be performed by anyone at any time who has both the instructions and the scoring system of the tests (Heale & Twycross, 2015). Tester two performing the testing for the first time and having less reliable scores, suggests that it is not possible for anyone to perform the testing successfully, even with a clinical background. Even though tester one was more familiar with the tests and had a higher reliability as a tester, tester one's internal consistency was slightly lower than tester twos. When having a closer look of the internal consistency, an acceptable value is anything between 0.70 to 0.95 with 0.9 being excellent (Tavakol & Dennick, 2011). Strength tests one, two and seven, had either poor ($0.6 > \alpha \geq 0.5$), unacceptable ($0.5 > \alpha$), or questionable ($0.7 > \alpha \geq 0.6$) internal consistency. Tests three,

four, five, eight and nine had the strongest scores with good to excellent internal consistency. Test six also showed high values, but still had more variance than the other five tests.

These results suggest that there are differences in the individual tests, which could mean the tests do not have clear enough instructions or scoring systems. Both the guidelines and the scoring system (Appendix 2) of the tests leave too much room for interpretation even if the two testers are therapists with a similar clinical background. The test instructions and guidelines between the tests vary, which can affect the scoring process. The scoring system could benefit from clearer differences between the scores and if that is not possible it could benefit from adding one more score. For strength tests to be reliable, they need to be repeatable as well. The simpler the test is to execute the less room it leaves for error.

When looking at the requirements of the sport, leg strength and power are important for the gymnasts in terms of performance (Dallas et al. 2021; Nitzsche et al. 2021; Behm et al. 2017; Alwasif, 2019; Lloyd & Oliver, 2012). Having high enough levels of leg strength is also part of injury prevention, since majority of the injuries are on the lower extremities (Alwasif, 2019; Thomas & Thomas, 2019; Dahab & McCambridge, 2009; Behm et al. 2017; Faigenbaum & Micheli, 2017). Tests two and seven highlight adductor and hamstring strength, but as tests they showed too low internal consistency to be seen as reliable strength tests. Tests three, eight and nine were the simplest to execute out of the nine strength tests, which could be one reason why they showed to be reliable. Field-based tests are widely recommended since they involve minimal equipment and have minimal costs which are easy to administer as well as allowing a larger number of gymnasts to be tested in a short period of time (Salse-Batán et al. 2022). In addition, clinical tests could be used as a part of the test battery by the right experts working in the field (Haitz et al. 2014; Barnett et al. 2015; Räisänen et al. 2016; Whatman, Hume & Hing, 2013).

Furthermore, it is important to consider do all the nine strength tests measure strength or are they more performance-based tests. Based on the theory alone, lower limb strength and power of athletes and aesthetic group gymnasts should be both increased and measured to help prevent the most common injuries in the sport. Jump tests and plyometric exercises focusing on leg strength are currently the most popular tests and movements to use with gymnasts. (Gasparetto

et al. 2022; Salse-Batán et al. 2022; Lindberg et al. 2022; Dallas et al. 2021; Deley et al. 2010; Nitzsche et al. 2021)

Individual tests showed a lot of variation in terms of the inter- and intra-rater reliability. Six out of nine strength tests had high internal consistency values, and therefore can be deemed as more reliable strength tests. Based on the results and the research, sport specificity is higher in only 3 out of 9 tests. When looking at the test more closely as a total score per athlete, it showed promising results in terms of associating the athletes either as "stronger" or "weaker". Therefore, it was important to see if there is a way for the coaches to use the existing strength tests as a battery by summing up the overall results per athlete (max 36 points). Because a high level of strength is needed in the sport to both perform at the required level as well as to prevent injuries, it is important for coaches to be able to identify the gymnasts who need strength early on. This creates the importance of measuring strength throughout the training life cycle. (Sawczyn et al. 2016; Dallas et al. 2021; Alwasif, 2019; Thomas & Thomas, 2019; Gasparetto, 2022; Salse-Batán et al. 2022; Osmala et al. 2021)

In the field, there are currently no validated tools for strength testing purposes, especially with good validity and reliability (Salse-Batán et al. 2022). Based on the results, this research offers a simple and reliable tool for coaches or for anyone to use to identify these much needed qualities in athletes which are important already in the early stages of the training.

# 8 Strengths, weaknesses, and reliability of the study

This study has strengths and weaknesses. To achieve more definitive results, a third tester and a third testing time could be recommended. However, to achieve more accurate results the tests could be modified to have simpler and clearer instructions and a larger population could be tested, for example twenty gymnasts instead of the current nine gymnasts.

This test-retest was executed in a training camp environment, which had its pros and cons. It could be beneficial to not be testing in the same room as the gymnasts are training. The gymnasts were already warm by the ongoing training, but on the first testing day one of the coaches started to rehearse the strength test movements with material of their own. The situation needed to be quickly addressed which interrupted the testing session of one of the gymnasts. It was important for the reliability of the test-retest that tests were not rehearsed beforehand other than as a part of the testing. If the tests are rehearsed before the testing situation, the gymnasts might be more tired, which could lead to a greater risk of injury, as well as instructions given before by a coach might contradict with the actual testing protocol for this study. It was important that the tests were introduced to the gymnasts first time in the start of each test on the first testing day to ensure the integrity of the testing protocol.

On the second day, it was clear to see that the gymnasts were more tired compared to the first testing day even though the reliability of the scores were higher. The difference in scores could also suggest that the testers and gymnasts were more familiar with the tests and therefore performed more consistently. After interviewing the gymnasts separately, it became clear that they had been training and rehearsing the tests between the testing days by the coaches as well as individually. Tiredness could be one of the reasons that led to lower scores on the second testing day.

In addition, on the second testing day, music was played loudly by an individual coach which affected some of the gymnast's concentration during the testing. Prior to this moment no music had been played on day one. The eighth test requires a metronome to help with the right pace of the heel raises, which the athlete found hard to hear properly due to the music. Going further with the research it would be beneficial to have a separate space for the testing as well as a longer time frame between the testing sessions.

The strengths of this test-retest are the clear testing protocol and systematic approach to the practical side of the research being done. Also, it is a benefit that the same number of gymnasts was tested on both testing days. The ethical principles were met with high standards and the ethical approval was applied long before the start of the testing. Scoring of the tests being single blinded by two qualified physiotherapists also decreases the level of bias. In addition, an evidence-based practical application was created based on this test-retest. Even though the testing sessions were identical in terms of time and location, testing days being only two days apart, might have affected the gymnasts scores negatively. If the gymnasts were given more rest between the testing days and the test movements would not have been rehearsed, it could have impacted the results positively leading into more reliable results overall. Still, the difficulties with the grading of the tests with the current guidelines further impacted the scoring and would need to be addressed before going further.

Based on the researcher's background as a physiotherapist, simpler tests would also be recommended. Clinical tests are widely used when assessing patients and athletes' performance as well as inexpensive to use, therefore could be a helpful addition to the test battery. In addition, guidelines for the scoring and execution needs to be more defined. A clearer scoring system would leave less room for interpretation and is therefore highly recommended.

Overall, this study has many strengths. Even though the testing days included some minor incidences in the testing sessions, the test-retest went seamlessly from start to finish and created a strong basis for the reliable evidence stated in this study.

# 9 Conclusions and development proposals

To conclude, competing gymnasts have a high risk of injury due to the large amount of training hours and the demands of the sport. Female gymnasts have a high injury incidence of 9.37 per 1000 athlete exposures. In AGG sufficient strength characteristics support athlete's performance and prevents injuries, especially among young athletes. There are no current such testing batteries available, let alone evidence-based testing batteries. Therefore, in AGG there is a need for meaningful, reliable, and sensitive outcome gymnastic-specific fitness tests. This is the first time such a project has been attempted, which is an important and essential first step in creating a final testing protocol for AGG and strength.

Based on the reliable results found in this study, the athlete's total scores of the current strength tests can be used to identify "stronger" and "weaker" athletes. An athlete identification tool was created to identify stronger to weaker athletes. This tool is simple to use, easy to implement in the field and is designed to benefit the athlete. The tool will enable a coach to individualise the training to meet the demands of the sport for athletes to perform better as well as prevent future injuries.

Furthermore, clearer guidelines and testing protocols for the strength tests are recommended and more distinctive differences between the scores of the scoring system are needed. Furthermore, another recommendation would be for the testers to practice the strength tests before using them to increase the reliability. Tests assessing leg strength and power are beneficial for the performance and gymnastic skills as well as a tool for injury prevention in AGG. Therefore, tests assessing leg strength and power should be added to the future test battery. Strength tests three, eight, and nine have proven to be reliable for AGG, therefore should be kept in the test battery in the future as they have now proven to be evidence-based strength tests. Overall, this is an ethical high-quality study that offers evidence-based findings to benefit and support today’s athletes in gymnastics.

## References

Alwasif, N. O. (2019). Specific strength training on parallel bars and its influence on the technical performance level of Gymnastics. Turnen trainieren und vermitteln: 10. Jahrestagung der dvs-Kommission Gerätturnen Conference 2018, Göttingen, Germany. https://www.researchgate.net/publication/333352702_Specific_strength_training_on_parallel_bars_and_its_influence_on_the_technical_performance_level_of_Gymnastics

Barnett, L., Reynolds, J., Faigenbaum, A. D., Smith, J. J., Harries, S., & Lubans, D. R. (2015). Rater agreement of a test battery designed to assess adolescents' resistance training skill competency. *Journal of Science and Medicine in Sport*, 18(1), 72–76. https://doi.org/10.1016/j.jsams.2013.11.012

Behm, D. G., Young, J. D., Whitten, J. H. D., Reid, J. C., Quigley, P. J., Low, J., Li, Y., Lima, C. D., Hodgson, D. D., Chaouachi, A., Prieske, O., & Granacher, U. (2017). Effectiveness of Traditional Strength vs. Power Training on Muscle Strength, Power, and Speed with Youth: A Systematic Review and Meta-Analysis. *Frontiers in Physiology*, 8, 423. https://doi.org/10.3389/fphys.2017.00423

Bobak, C., Barr, P. & O'Malley, A. (2018). Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. *BMC Medical Research Methodology*, (18), 93. https://doi.org/10.1186/s12874-018-0550-6

Campbell, R. A., Bradshaw, E. J., Ball, N. B., Pease, D. L., & Spratford, W. (2019). Injury epidemiology and risk factors in competitive artistic gymnasts: a systematic review. *British Journal of Sports Medicine*, (53), 1056-1069. https://doi.org/10.1136/bjsports-2018-099547

Collins, L. M. (2007). Research Design and Methods. In J. E. Birren (Ed.), *Encyclopedia of Gerontology* (2nd ed.), (pp. 433-442). Elsevier. https://doi.org/10.1016/B0-12-370870-2/00162-1

Dahab, K. S., & McCambridge, T. M. (2009). Strength training in children and adolescents: raising the bar for young athletes? *Sports Health*, 1(3), 223–226. https://doi.org/10.1177/1941738109334215

Dallas, G. C., Dallas, C. & Maridaki, M. (2021). The effect of 10-week isokinetic training on muscle strength and gymnastic performance in preadolescent female gymnast. *Science of Gymnastics* Journal, 13*(3), 399-409*. https://doi.org/10.52165/sgj.13.3.399-409

Deley, G., Cometti, C., Fatnassi, A., Paizis, C., & Babault, N. (2011). Effects of combined electromyostimulation and gymnastics training in prepubertal girls. *Journal of strength and conditioning research*, *25*(2), 520–526. https://doi.org/10.1519/JSC.0b013e3181bac451

Faigenbaum, A. Chu, D. (2017). Plyometric Training for Children and Adolescents. *American College of Sports Medicine*. https://www.acsm.org/docs/default-source/files-for-resource-library/smb-plyometric-training-for-children-and-adolescents.pdf?sfvrsn=fcc67055_2

Faigenbaum, A. D., & McFarland, J. E. (2016). RESISTANCE TRAINING FOR KIDS. *Acsm's Health & Fitness Journal*, 20(5), 16–22. https://doi.org/10.1249/fit.0000000000000236

Faigenbaum, A. & Micheli, L. (2017). Youth Strength Training. ACSM Sports Medicine Basics. *American College of Sports Medicine*. https://www.acsm.org/docs/default-source/files-for-resource-library/smb-youth-strength-training.pdf?sfvrsn=85a44429_2

Faigenbaum, A., Kraemer, W., Blimkie, C. J., Jeffreys, I., Micheli, L., Nitka, M., Rowland, T, (2009). Youth Resistance Training: Updated Position Statement Paper from the National Strength and Conditioning Association. *Journal of Strength and Conditioning Research*, (23), 60-79. https://doi.org/10.1519/JSC.0b013e31819df407

Fairman, C. M., LaFountain, R. L., Lucas, A. R., Focht, B. C. (2018). Monitoring Resistance Exercise Intensity Using Ratings of Perceived Exertion in Previously Untrained Patients with Prostate Cancer Undergoing Androgen Deprivation Therapy. *Journal of Strength and Conditioning Research*, 32(5), 1360-1365. https://doi.org/10.1519/JSC.0000000000001991

Fernando, J. (2023). R-Squared: Definition, Calculation Formula, Uses, and Limitations. Investopedia. https://www.investopedia.com/terms/r/r-squared.asp

Fink, A. (2010). Survey Research Methods. In P. Peterson, E. Baker & B. McGaw (Eds.), *International Encyclopedia of Education (3rd ed.)*, (pp. 152-160). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00296-7

Gasparetto, Z., Julião, A. L., Thuany, M., Martinez, P. F., de Mendonça Bacciotti, S., & de Oliveira-Junior, S. A. (2022). Concerns about strength tests in gymnastics: a systematic review. *Science of Gymnastics Journal*, 14(2), 225–236. https://doi.org/10.52165/sgj.14.2.225-236

Giavarina D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, 25(2), 141–151. https://doi.org/10.11613/BM.2015.015

Grgic, J., Lazinica, B., Schoenfeld, B.J. & Pedisic, Z. (2020). Test–Retest Reliability of the One-Repetition Maximum (1RM) Strength Assessment: A Systematic Review. *Sports Medicine – Open*, 6(31). https://doi.org/10.1186/s40798-020-00260-z

Guyatt, G., H., Oxman, A. D., Kunz, R., Vist, G. E., Falck-Ytter, Y., Schünemann, H. J. (2008). What is "quality of evidence" and why is it important to clinicians? *British Medical Journal*, 336, 995. https://doi.org/10.1136/bmj.39490.551019.BE

Haitz, K., Shultz, R., Hodgins, M., & Matheson, G. O. (2014). Test-retest and interrater reliability of the functional lower extremity evaluation. *The Journal of orthopaedic and sports physical therapy*, *44*(12), 947–954. https://doi.org/10.2519/jospt.2014.4809

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-based nursing*, 18(3), 66–67. https://doi.org/10.1136/eb-2015-102129

Hébert-Losier, K., Wessman, C., Alricsson, M., & Svantesson, U. (2017). Updated reliability and normative values for the standing heel-rise test in healthy adults. *Physiotherapy Journal*, 103(4), 446–452. https://doi.org/10.1016/j.physio.2017.03.002

Kolar, E., Pavletič, M. S., Smrdu, M. & Atiković, A. (2017). Athletes' perception of the causes of injury in gymnastics. *The Journal of Sports Medicine and Physical Fitness*, (57), 703-10. https://pubmed.ncbi.nlm.nih.gov/27029958/

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kraemer, W. & Ratamess, N. A. (2004). Fundamentals of Resistance Training: Progression and Exercise Prescription. *Medicine & Science in Sports & Exercise*, 36(4), 674-688. https://doi.org/10.1249/01.MSS.0000121945.36635.61

Lindberg, K., Solberg, P., Bjørnsen, T., Helland, C., Rønnestad, B., Thorsen Frank, M., Haugen, T., Østerås, S., Kristoffersen, M., Midttun, M., Sæland, F., Eythorsdottir, I., & Paulsen, G. (2022). Strength and Power Testing of Athletes: A Multicenter Study of Test–Retest Reliability. *International Journal of Sports Physiology and Performance*, 17(7), 1103-1110. https://doi.org/10.1123/ijspp.2021-0558

Lloyd, R. S., Faigenbaum, A. D., Stone, M. H., Oliver, J. L., Jeffreys, I., Moody, J. A., Brewer, C., Pierce, K. C., McCambridge, T. M., Howard, R., Herrington, L., Hainline, B., Micheli, L. J., Jaques, R., Kraemer, W. J., McBride, M. G., Best, T. M., Chu, D. A., Alvar, B. A. & Myer, G. D. (2014). Position statement on youth resistance training: the 2014 International Consensus. *British Journal of Sports Medicine*, 48 (7), 498–505. https://pubmed.ncbi.nlm.nih.gov/24055781/

Lloyd, R. S. & Oliver, J. L. (2012). The Youth Physical Development Model: A New Approach to Long-Term Athletic Development. *Strength and Conditioning Journal,* 34(3), 61-72. https://doi.org/10.1519/SSC.0b013e31825760ea

Martínez, P.T., & Grande, I. (2021). Analysis and comparison of training load between two groups of women's artistic gymnasts related to the perception of effort and the rating of the perceived effort session. *Science of Gymnastics Journal.* https://doi.org/10.52165/sgj.13.1.19-33

Marx, R. G., Menezes, A., Horovitz, L., Jones, E. C., & Warren, R. F. (2003). A comparison of two-time intervals for test-retest reliability of health status instruments. *Journal of Clinical Epidemiology*, 56(8), 730–735. https://doi.org/10.1016/s0895-4356(03)00084-2

Muñoz-Bermejo, L., Adsuar, J. C., Mendoza-Muñoz, M., Barrios-Fernández, S., Garcia-Gordillo, M. A., Pérez-Gómez, J., & Carlos-Vivas, J. (2021). Test-Retest Reliability of Five Times Sit to Stand Test (FTSST) in Adults: A Systematic Review and Meta-Analysis. *Biology*, 10(6), 510. https://doi.org/10.3390/biology10060510

Nitzsche, N., Siebert, T., Schulz, H., & Stutzig, N. (2021). Effect of plyometric training on dynamic leg strength and jumping performance in rhythmic gymnastics: A preliminary study. *Isokinetics and Exercise Science*. https://doi.org/10.3233/IES-210148

Osmala, J., Pitkänen, A. & Vastamäki, S. (2021). Liikkuvuusharjoittelu. Suomen Voimisteluliitto. https://www.voimistelu.fi/Portals/0/Lajit%20yleist%C3%A4/Dokumentit/Liikkuvuusharjoittelu%20-%20Voimisteluliitto%203-2021_FINAL.pdf

Pei, Y. A., Mahmoud, M. A., Baldwin, K., & Franklin, C. (2022). Comparing Musculoskeletal Injuries across Dance and Gymnastics in Adolescent Females Presenting to Emergency Departments. *International journal of environmental research and public health*, 20(1), 471. https://doi.org/10.3390/ijerph20010471

Reed, J. & Bowen, J. D. (2008). Principles of Sports Rehabilitation. In P. H. Seidenberg & A. I. Beutler (Eds.). *The Sports Medicine Resource Manual*, 33, 431-436. https://doi.org/10.1016/B978-141603197-0.10033-3

Räisänen, A., Pasanen, K., Krosshaug, T., Avela, J., Perttunen, J., & Parkkari, J. (2016). Single-Leg Squat as a Tool to Evaluate Young Athletes' Frontal Plane Knee Control. *Clinical journal of sport medicine: official journal of the Canadian Academy of Sport Medicine*, 26(6), 478–482. https://doi.org/10.1097/JSM.0000000000000288

Salse-Batán, J., Varela, S., García-Fresneda, A. & Ayán, C. (2022). Reliability and validity of field-based tests for assessing physical fitness in gymnasts. *Apunts Sports Medicine*, 57(216). https://doi.org/10.1016/j.apunsm.2022.100397

Sands, W. A., Wurth, J. J. & Hewit, J. K. (2012). The National Strength and Conditioning Association's (NSCA). *Basics of Strength and Conditioning Manual*, 9-10. https://www.nsca.com/contentassets/116c55d64e1343d2b264e05aaf158a91/basics_of_strength_and_conditioning_manual.pdf

Santalov, K. (n.d.). IFAGG. About AGG. International Federation of Aesthetic Group Gymnastics, 2004-2023. https://www.ifagg.com/v1/page.php?n=6

Sawczyn, S., Zasada, M., Kochanowicz, A., Niespodziński, B. Sawczyn, M. & Mishchenko, V. (2016). The effect of specific strength training on the quality of gymnastic elements execution in young gymnasts. *Baltic Journal of Health and Physical Activity*, 8(4), 79-91. https://doi.org/10.29359/BJHPA.08.4.09

Suchomel, T.J., Nimphius, S. & Stone, M.H. (2016). The Importance of Muscular Strength in Athletic Performance. *Sports Medicine*, 46, 1419–1449. https://doi.org/10.1007/s40279-016-0486-0

Sun, Z. (2023). Analyzing the effects of strength-trainingbased artistic gymasic teaching. *Rev Bras Med Esporte,* 29(1). https://doi.org/10.1590/1517-8692202329012022_0225

Suomen Voimisteluliitto. (N.d.). Introductions to the types of competitive gymnastics. Suomen Voimisteluliitto. https://www.voimistelu.fi/voimisteluliitto/in-english/introductions-to-the-types-of-competitive-gymnastics/

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tayech, A., Mejri, M. A., Chaabene, H., Chaouachi, M., Behm, D. G., & Chaouachi, A. (2019). Test-retest reliability and criterion validity of a new Taekwondo Anaerobic Intermittent Kick Test. *The Journal of sports medicine and physical fitness*, 59(2), 230–237. https://doi.org/10.23736/S0022-4707.18.08105-7

The Finnish National Board on Research Integrity (TENK). (2021). Guidelines for ethical review in human sciences. Finnish National Board on Research Integrity TENK. https://tenk.fi/en/advice-and-materials/guidelines-ethical-review-human-sciences

Thomas, R. E., & Thomas, B. C. (2019). A systematic review of injuries in gymnastics. *The Physician and Sports medicine,* (47), 96-121. https://doi.org/10.1080/00913847.2018.1527646

Whatman, C., Hume, P., & Hing, W. (2013). The reliability and validity of physiotherapist visual rating of dynamic pelvis and knee alignment in young athletes. *Physical therapy in sport: official journal of the Association of Chartered Physiotherapists in Sports Medicine*, *14*(3), 168–174. https://doi.org/10.1016/j.ptsp.2012.07.001

Xiao, M. (2023). Specificity and diversity of athletic ability needs among professional gymnasts. *Rev Bras Med Esporte,* 29(12). https://doi.org/10.1590/1517-8692202329012022_0275

Xu, M., Fralick, D., Zheng, J. Z., Wang, B., Tu, X. M., & Feng, C. (2017). The Differences and Similarities Between Two-Sample T-Test and Paired T-Test. *Shanghai archives of psychiatry,* 29(3), 184–188. https://doi.org/10.11919/j.issn.1002-0829.217070

Zhuang, L., Zhu, S., & Shi, Y. (2023). SYSTEMATIC SKILL PRACTICE IN WOMEN's FLOOR EXERCISE. *Rev Bras Med Esporte,* 29(1). https://doi.org/10.1590/1517-8692202329012022_0367

# Appendices

## Appendix 1. The theoretical basis

The PubMed database was used to attain high quality resources for the theoretical basis of this test-retest study. The search was successful, even though a lot of the material could not be found through the PubMed database. A grey literature search was also performed using search engines such as Google Scholar and ResearchGate to supplement the search.

The themes in the literature review are test-rest, strength, strength testing and gymnastics. To narrow the search down filters such as *10 years*, *full text*, *female*, and *human* were used. Filters were added to narrow the search down after search terms such as test and test-retest are widely used in research. The filter *human* was used only to exclude mice studies.

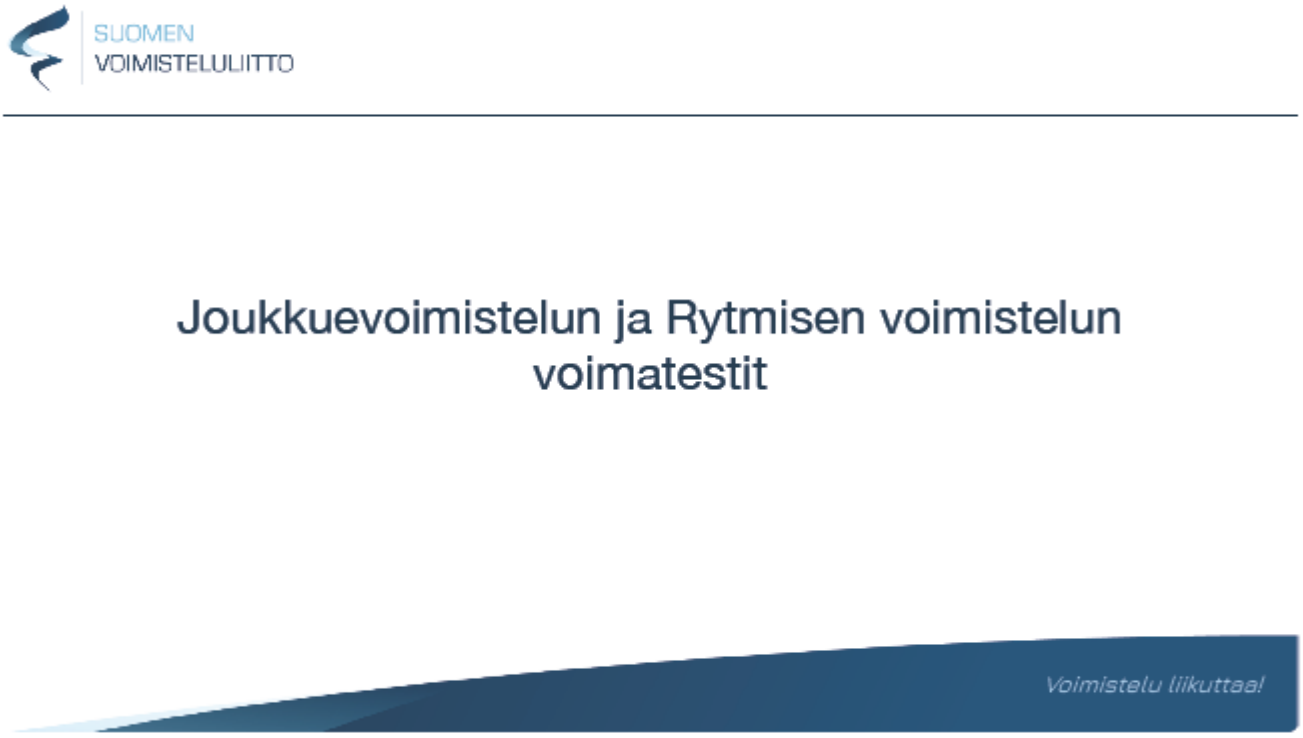
The search through the PubMed database did not show good quality studies highlighting muscle strength in gymnastics or strength testing in gymnastics. Many of the eligible studies found later, had missing key words or a low number of key words, which made the papers hard to find to begin with. In addition, some of the studies were completed over 10 years ago. In order to find more studies, complementary sources such as Google Scholar and ResearchGate were used with the same search terms and search strategy.

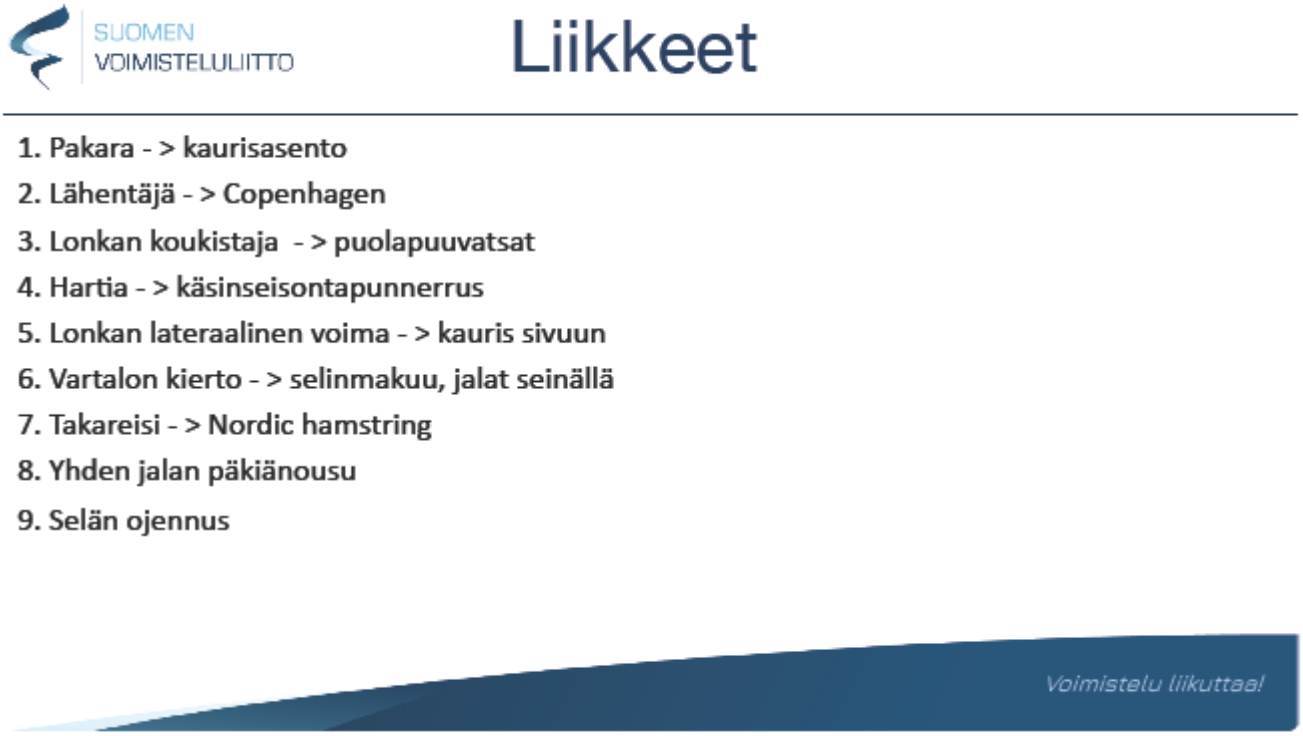
A total of 2134 articles was identified after excluding duplicates (n=34) and 280 articles were assessed for eligibility. After screening all the 280 articles, none of the studies found matched the themes of this research. After broadening the search further with a grey literature search, a total of 6 articles were included as shown in Figure 1. Table 11 states the search terms, Boolean operators, and the search strategy used for the theoretical basis of this study.

Table 11. The Pubmed database search

| **Searches in the PubMed Database** | **Results** |
|---|---|
| Gymnastics OR gymnast OR artistic gymnast OR aesthetic gymnast OR aesthetic group gymnast OR AGG OR rhythmic gymnast | 7,179 |
| Aesthetic sport OR artistic sport OR dance sport OR sport OR youth sport | 429,396 |
| Dance* OR dancer OR dancing OR dance athlete OR athlete OR young athlete OR elite | 414,829 |
| 1 OR 2 OR 3 | 475,130 |
| Test retest* OR test-retest | 38,884 |
| strength test* OR performance test* OR sport test* OR movement test* OR field test* | 1,484,318 |
| Reliability test OR validity test | 342,875 |
| 5 OR 6 OR 7 | 1,712,240 |
| Intrarater OR interrater OR inter- AND intra-rater OR rater reliability | 19,124 |
| reliability OR validity OR repeatability | 607,441 |
| 9 OR 10 | 608,086 |
| Strength* OR muscle strength OR power OR performance OR fitness OR exercise | 5,439,501 |
| 4 AND 8 AND 11 | 10,751 |
| 4 AND 8 AND 11 AND 12 | 8,428 |
| Filter: 10 years | 5,451 |
| Filter: 10 years, Full text | 5,387 |
| Filter: 10 years, Full text, Female | 2,226 |
| Filter: 10 years, Full text, Female, Human | **2,194** |

**Appendix 2. The Finnish Gymnastics Federation's strength tests**

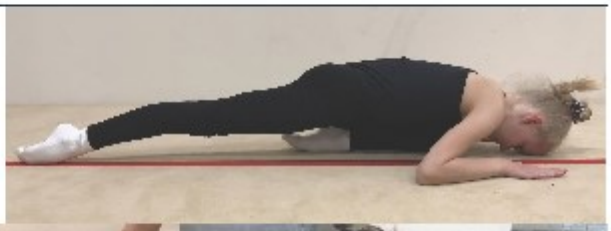


SUOMEN VOIMISTELULIITTO

# 1.Pakara - kaurisasento

- Testattava asettuu viivan päälle kaurisasentoon siten, että suoran jalan nilkka, polvi lonkka ja hartia ovat viivan päällä
- Suoran jalan lonkka ei saa lähteä kääntymään auki, takajalan polvi osoittaa kohti lattiaa
- Koukkujalan kantapää tarkasti pakaran alla
- Otsa lattiassa, katse lattiaan
- Kämmenet maassa otsan tasolla, kyynärpäät maassa vartalonsuuntaisesti
- Polvi ilmassa lähtöasennossa nilkka ojennettuna
- Avustaja asettaa kädet lantiolle kevyesti merkiksi, tavoitteena havaita lantion mahdollinen kiertyminen tai nouseminen



**Asteikko**

- 4 pistettä= otsa pysyy maassa, pakara pysyy kiinni kantapäässä, jalka nousee polvi suorana hallitusti ylös lattiasta, jalka pysyy testiviivan yläpuolella vaakatasossa paikallaan noin 5 sekunnin ajan
- 2 pistettä= otsa pysyy maassa, pakara pysyy kiinni kantapäässä, jalka nousee alle vaakatason polvi suorana ilmassa, mutta ei pysy noin 5 sekuntia, tärisee tai ei pysy testiviivan yläpuolella
- 0 pistettä= ei pääse testiasentoon, otsa irtoaa maasta tai pakara kantapäästä, jalka ei nouse lattiasta

*Voimistelu liikuttaa!*

SUOMEN VOIMISTELULIITTO

# 2.Lonkan lähentäjät

- Testattava asettuu viivan päälle, alajalan nilkka, polvi, lantio ja kyynärpää viivan päällä
- Ylemmän nilkan kehräsluu penkin päälle, jalka suoraa alajalan yläpuolella
- Vapaa käsi vyötäröllä
- Testattava nostaa itsensä kyljkilankkuun lonkat nollakulmassa, alajalka vielä maassa, molemmat pakarat osuvat takana olevaan keppiin, vartalon neliö osoittaa samaan suuntaan
- Testattava nostaa alajalan ilmaan vasta kun testaaja antaa siihen luvan

**Asteikko**

- 4 pistettä = asento pysyy hallittuna 5 sekunnin ajan, alajalan kehräsluun irtoamatta penkistä
- 2 pistettä= keskivartalon asento säilyy hallittuna keinumatta puolelta toiselle, jalka pysyy ilmassa 5 sekunnin ajan suoraan toisen jalan alla, mutta ei pysy penkissä kiinni
- 0 pistettä= ei pääse testiasentoon jalka maassa, keskivartalon hallinta ei säily, tai jalka ei irtoa lattiasta/pysy ilmassa 5 sekuntia

*Voimistelu liikuttaa!*

SUOMEN VOIMISTELULIITTO

# 3.Lonkan koukistajat / Alavatsa

- Testattava roikkuu lähtöasennossa puolapuilla, peukalot hartialeveydellä, myötäote ja jalat puristettuna yhteen
- Testattava lähtee nostamaan suoria ja yhteen puristettuja jalkoja ylös, tarkoituksena osua jaloilla puolaan tai otsaan
- Sarja täytyy toistaa ilman taukoja

**Asteikko**

- 4 pistettä= 10 nostoa
- 2 pistettä= 5 nostoa
- 0 pistettä= ei pääse testiasentoon tai saa vähemmän kuin 5 onnistunutta nostoa

*Voimistelu liikuttaa!*

SUOMEN VOIMISTELULIITTO

# 4.Hartia - käsinseisontapunnerrus

- Testattava asettaa kädet maahan kahden jalanmitan päähän seinästä, sormet osoittaen eteenpäin ja nousee käsinseisontaan rintamasuunta seinää kohden
- Jalat yhdessä tai enintään hartialevyisessä haarassa
- Testattavan tulee pystyä pysäyttämään liike missä tahansa vaiheessa suoritusta eli niin sanottua vapaapudotusta ei sallita

**Asteikko**

- 4 pistettä= hallittu alasmeno, pää hipaisee alustaa (paino ei siirry pään varaan) hallittu nosto ylös takaisin lähtöasentoon
- 2 pistettä=hallittu alasmeno, mutta ei pääse ylös, paino saa mennä pään päälle
- 0 pistettä= ei pääse testiasentoon tai liike ei onnistu

*Voimistelu liikuttaa!*

SUOMEN VOIMISTELULIITTO

# 5.Lonkan lateraalinen voima

- Testattava asettuu kilpikonna-asentoon takajalka 45-60 asteen kulmassa takaviistossa, alla olevan jalan kantapää tarkasti pakaran alla
- Otsa kiinni kämmenselässä, katse lattiaan
- Kämmenet maassa otsan alla, kyynärpäät maassa sivulle osoittaen
- Jalan saa nostaa koukussa tai suorana, suoran jalan lähdössä polvi loppuun asti ojennettuna, koukussa takajalan polvi on 90 asteen kulmassa
- Koukkujalkaa nostaessa takajalan nilkan ja polven täytyy säilyä noston ajan samassa tasossa
- Testattavan jalan puoleinen lonkka ei saa lähteä kääntymään auki nostossa
- Avustaja asettaa kädet kevyesti merkiksi lantiolle, tavoitteena havaita lantion mahdollinen kiertyminen tai nouseminen.
- Lonkan kulman täytyy pysyä samana liikkeen ajan

**Asteikko**

- 4 pistettä=otsa pysyy kiinni kämmenselissä, pakara pysyy kiinni kantapäässä, jalka nousee polvi suorana hallitusti ylös lattiasta ja pysyy noin 5 sekunnin ajan ilmassa paikallaan
- 2 pistetttä=otsa pysyy kämmenselissä, pakara pysyy kiinni kantapäässä, jalka nousee koukussa ilmaan, nilkka ja polvi samalla tasolla, asento säilyy 5 sekuntia
- 0 pistettä= ei pääse testiasentoon, otsa irtoaa maasta tai pakara kantapäästä, jalka ei nouse lattiasta

*Voimistelu liikuttaa!*

SUOMEN VOIMISTELULIITTO

# 6.Vartalon kierto

- Lähtö jalat sivulta
- Vartalo 90 asteen kulmassa, lonkat päällekkäin, molemmat pakarat kiinni seinässä
- Jalkojen puoleinen käsi vaakatasossa sivulla ja toinen käsi kohti kattoa hartia maassa
- Jalkojen nosto hallitusti jalat yhdessä kohti kattoa ja hallitusti takaisin alas
- Hartia ei irtoa lattiasta
- Jalkojen on pysyttävä yhdessä polvet suorina koko suorituksen ajan

**Asteikko**

- 4 pistettä= hartia pysyy lattiassa, jalkojen nosto hallitusti yhdessä ylös ja takaisin alas
- 2 pistettä= hartia pysyy maassa, jalkojen nosto ylös yhdessä hallitusti, alastulo ei onnistu hallitusti
- 0 pistettä= ei pääse testiasentoon, hartia irtoaa lattiasta, jalat eivät nouse lattiasta tai jalat aukeavat nostossa

*Voimistelu liikuttaa!*

SUOMEN VOIMISTELULIITTO

# 7.Takareisi - Nordic hamstring

- Kädet suorana alaviistossa (noin 45 asteen kulmassa)
- Pari pitää jaloista kiinni, lantion levyinen asento
- Vartalo pysyy liikkumattomana suorituksen ajan
- Onnistuneessa suorituksessa testattava pääsee hallitusti alas ja ylös

**Asteikko**

- 4= Onnistuu alas ja ylös lonkat 0 kulmassa, keskisormet koskettaa lattiaa, paino ei kuitenkaan mene käsien päälle
- 2=Pääsee hallitusti alas, mutta ei ylös
- 0= Ei pääse testiasentoon, 0 kulma lonkissa ei säily alas mennessä, asento tippuu heti käsien varaan alhaalla

*Voimistelu liikuttaa!*

SUOMEN VOIMISTELULIITTO

# 8. Yhden jalan päkiänousu

- Testattava asettuu seisomaan seinän eteen, kevyt ote seinästä sormenpäillä - seinään ei saa nojata
- Mitataan testattavan jalan maksimi päkiänousukorkeus mittakepin avulla, asettamalla mittakeppi kantapään taakse
- Metronomi asetetaan 60bpm
- Päkiänousu metronomin tahtiin, 1s ylös ja 1 s alas
- Nousu joka kerta maksimiin
- Avustajan käsi on kevyesti testattavan polvessa, tarkistamassa että polvi pysyy suorana koko suorituksen ajan

**Asteikko**

- 4 pistettä = >30
- 2 pistettä = >25
- 0 pistettä = <25

*Voimistelu liikuttaa!*

SUOMEN VOIMISTELULIITTO

# 9. Selän ojennus

- Testattava asettuu päinmakuulle liukuasentoon, toinen keppi asetetaan testattavan niskan ja käsien väliin
- Kädet hartialeveydellä ja kyynärpäät suorana
- Jalat lukittuna yhdessä, pari istuu nilkkojen päällä
- Mittauskeppi on haarustan ja pakaran alareunan tasolla vartalon vieressä, sen verran irti vartalosta, etteivät testattavan kädet osu keppiin noston aikana
- Rauhallinen nosto ja noin 2 sekunnin pito

**Asteikko**

- 4 pistettä= Hartiakeppi osuu mittauskeppiin, 2 sekunnin pito
- 2 pistettä= Kädet ylittävät mittauskepin sivulta katsottuna, 2 sekunnin pito
- 0 pistettä= Ei pääse testiasentoon tai nosto ei onnistu riittävän korkealle

*Voimistelu liikuttaa!*

Syötä kuvateksti kirjoittamalla.

## Appendix 3. Warmup and Borg CR10 RPE scale

Table 12. Warmup

| Lämmittely (warmup) | 10-15min |
|---|---|
| 1. Hölkkä (jogging) | 3min |
| 2. Mittarimato (inch worm) | 10x |
| 3. Kyykky kehonpainolla (squat) | 10x |
| 4. Selkälihasliike liukuasennossa vatsamakuulla (back exercise in prone position) | 10x |
| 5. Punnerus polvet lattiassa (push up with knees on the floor) | 10x |
| 6. Vatsalihasliike selinmakuulla (core exercise in supine position) | 10x |
| 7. Askelkyykky paikalla molemmat puolet (lunge, both sides) | 10x |

Table 13. Borg CR10 RPE scale (Fairman et al. 2018)

| Borg CR10 RPE-taulukko | |
|---|---|
| Kuormittuneisuus | Kuvaus |
| 0 | Ei lainkaan |
| 0,5 | Erittäin heikko |
| 1 | Hyvin heikko |
| 2 | Heikko |
| 3 | Kohtalainen |
| 4 | Melko voimakas |
| 5 | Voimakas |
| 6 | - |
| **7** | **Hyvin voimakas** |
| 8 | - |
| 9 | - |
| 10 | Erittäin voimakas |

# Appendix 4. Intraclass correlation coefficients for athlete's total scores

**Intraclass Correlation Coefficient**

| | Intraclass Correlation[b] | 95% Confidence Interval | | F Test with True Value 0 | | | |
|---|---|---|---|---|---|---|---|
| | | Lower Bound | Upper Bound | Value | df1 | df2 | Sig |
| Single Measures | ,889[a] | ,579 | ,974 | 15,250 | 8 | 8 | <,001 |
| Average Measures | ,941[c] | ,734 | ,987 | 15,250 | 8 | 8 | <,001 |

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Figure 17. ICC of tester one's total scores

**Intraclass Correlation Coefficient**

| | Intraclass Correlation[b] | 95% Confidence Interval | | F Test with True Value 0 | | | |
|---|---|---|---|---|---|---|---|
| | | Lower Bound | Upper Bound | Value | df1 | df2 | Sig |
| Single Measures | ,855[a] | ,387 | ,967 | 18,500 | 8 | 8 | <,001 |
| Average Measures | ,922[c] | ,558 | ,983 | 18,500 | 8 | 8 | <,001 |

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Figure 18. ICC of tester two's total scores

**Intraclass Correlation Coefficient**

| | Intraclass Correlation[b] | 95% Confidence Interval | | F Test with True Value 0 | | | |
|---|---|---|---|---|---|---|---|
| | | Lower Bound | Upper Bound | Value | df1 | df2 | Sig |
| Single Measures | ,891[a] | ,364 | ,977 | 30,000 | 8 | 8 | <,001 |
| Average Measures | ,942[c] | ,533 | ,989 | 30,000 | 8 | 8 | <,001 |

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Figure 19. ICC of the scores on testing day one

**Intraclass Correlation Coefficient**

| | Intraclass Correlation[b] | 95% Confidence Interval | | F Test with True Value 0 | | | |
|---|---|---|---|---|---|---|---|
| | | Lower Bound | Upper Bound | Value | df1 | df2 | Sig |
| Single Measures | ,957[a] | ,825 | ,990 | 41,350 | 8 | 8 | <,001 |
| Average Measures | ,978[c] | ,904 | ,995 | 41,350 | 8 | 8 | <,001 |

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Figure 20. ICC of the scores on testing day two