Received: 13 April 2023 | Revised: 2 July 2023 | Accepted: 1 August 2023

DOI: 10.1111/sms.14473

ORIGINAL ARTICLE

WILEY

Test–retest reliability of ski-specific aerobic, sprint, and neuromuscular performance tests in highly trained cross-country skiers

Elias Bucher[1,2] | **Grégoire P. Millet**[2] | **Jon P. Wehrlin**[1] | **Thomas Steiner**[1]

[1]Section for Elite Sport, Swiss Federal Institute of Sport Magglingen, Magglingen, Switzerland

[2]Institute of Sport Sciences, University of Lausanne, Lausanne, Switzerland

Correspondence
Elias Bucher, Section for Elite Sport, Swiss Federal Institute of Sport Magglingen SFISM, Hohmattstrasse 2, Magglingen CH-2532, Switzerland.
Email: elias.bucher@baspo.admin.ch

Funding information
The Swiss Federal Institute of Sport, Magglingen

Abstract

Purpose: Laboratory tests are commonly performed by cross-country (XC) skiers due to the challenges of obtaining reliable performance indicators on snow. However, only a few studies have reported reliability data for ski-specific test protocols. Therefore, this study examined the test–retest reliability of ski-specific aerobic, sprint, and neuromuscular performance tests.

Methods: Thirty-nine highly trained XC skiers (26 men and 13 women, age: 22 ± 4 years, $\dot{V}O_{2max}$: 70.1 ± 4.5 and 58.8 ± 4.4 mL·kg^{-1}·min^{-1}, respectively) performed two test trials within 6 days of a diagonal $\dot{V}O_{2max}$ test, $n = 27$; skating graded exercise test to assess the second lactate threshold (LT_2), $n = 27$; 24-min double poling time trial (24-min DP, $n = 25$), double poling sprint test ($Sprint_{DP1}$, $n = 27$), and 1-min self-paced skating sprint test ($Sprint_{1\text{-}min}$, $n = 26$) using roller skis on a treadmill, and an upper-body strength test (UB-ST, $n = 27$) to assess peak power (P_{peak}) with light, medium, and heavy loads. For each test, the coefficient of variation (CV), intraclass correlation coefficient (ICC), and minimal detectable change (MDC) were calculated.

Results: $\dot{V}O_{2max}$ demonstrated good-to-excellent reliability (CV = 1.4%; ICC = 0.99; MDC = 112 mL·min^{-1}), whereas moderate-to-excellent reliability was found for LT_2 (CV = 3.1%; ICC = 0.95). Performance during 24-min DP, $Sprint_{DP1}$, and $Sprint_{1\text{-}min}$ showed good-to-excellent reliability (CV = 1.0%–2.3%; ICC = 0.96–0.99). Absolute reliability for UB-ST P_{peak} was poor (CV = 4.9%–7.8%), while relative reliability was excellent (ICC = 0.93–0.97) across the loads.

Conclusion: In highly trained XC skiers, sport-specific aerobic and sprint performance tests demonstrated high test–retest reliability, while neuromuscular performance for the upper body was less reliable. Using the presented protocols, practitioners can assess within- and between-season changes in relevant performance indicators.

Jon P. Wehrlin and Thomas Steiner should be considered joint senior authors.

KEYWORDS
biathlon, Douglas bag, maximum oxygen uptake, Nordic ski, physiological testing, reproducibility

1 | INTRODUCTION

Laboratory physiological testing has a long history in elite cross-country (XC) skiing and has become a standard feature to support the scientific training approach.[1] Testing allows athletes, coaches, and sport scientists to objectively assess the effects of training interventions, track progress, and identify athletic potential.[2] Quantifying XC skiing performance in the field remains a challenge because skiing velocity on snow and roller skis can fluctuate between sessions on the same track due to the variations in terrain topography and the ever-changing friction resistances influenced by changing snow conditions, pavement surface, and temperatures. Therefore, accurate real-time information on power output is currently lacking to complement the intensity feedback of the physiological responses available during skiing. As a result, elite XC skiers typically undergo several standardized tests during their season preparation to measure key performance indicators, evaluate training load responses, and establish intensity zones for optimal training prescription. The key performance indicators in XC skiing include maximum oxygen consumption ($\dot{V}O_{2max}$),[3–7] performance at the second lactate threshold,[3,4] and gross efficiency.[5] As skiing speeds during competitions have increased substantially during the last few decades, and competitions are often decided during the final sprint, anaerobic capacity,[6] maximum skiing speed,[4,5,8] upper-body strength, and power capacity[7,9] further complement the performance framework in modern sprint and distance XC skiing.

Elite athletes demonstrate relatively small changes in laboratory-based performance over a season, highlighting the need for accurate measurements. Between the early preparation and competition periods, a non-significant change of 1.3 ± 2.4% in peak oxygen uptake ($\dot{V}O_{2peak}$) has been observed in elite XC skiers, with more variation in 1000-m time-trial performance (−7.4 ± 1.9%) and accumulated oxygen deficit ($\sum O_2$-deficit) (24.0 ± 19.5%) in the same period.[10] Furthermore, within- and between-season competition performance variability in male and female World Cup XC skiers is as small as 1.5%–2.2%, with an even smaller variability of 1.1%–1.5% in the top 10 ranked skiers.[11] Therefore, testing must be reliable to distinguish between meaningful changes, trends, and inherent test variability.[12,13] Test reliability depends on many factors, including measurement error, test protocol, exercise mode, standardization, equipment, athlete's performance level, motivation, and day-to-day biological variation.[2,13] Despite the extensive use of performance testing in XC skiing, reports on the test–retest reliability of ski-specific test protocols to assess aerobic, sprint or neuromuscular capacity are limited.

To our knowledge, only five studies have reported reliability data for XC ski-specific aerobic performance indicators. Losnegard et al.[10,14] have reported coefficient of variation (CV) values of 2.3% and 3.6%, respectively, for $\dot{V}O_{2peak}$ assessed during 1000-m and 800-m skating time trials on the treadmill in elite skiers. The CV for power output in well-trained XC skiers using custom-made and commercially available ski ergometers has been reported to range between 1.4% and 2.3% for incremental double poling (DP) protocols[15,16] and 3.0% for the 60-s DP test.[17] The same studies determined reliability values for $\dot{V}O_{2peak}$ between 1.9% and 2.5% during incremental[15,16] and 6-min constant[17] DP tests. However, it is recommended to measure $\dot{V}O_{2max}$ in XC skiers during an exercise mode involving a large muscle mass,[18] such as diagonal skiing on the treadmill. Reliability data for tests specifically designed to elicit $\dot{V}O_{2max}$ are currently limited to non-specific test forms (e.g., ski striding on foot) or tests on skiing ergometers. Four studies investigated the reliability of ski-specific sprint, strength, and power determinants. Stöggl et al.[8] and Losnegard et al.[10,14] reported the reliability of treadmill-based sprint-specific tests in elite skiers. Test–retest reliability of a DP maximum speed test (V_{max}: CV = 1.7%) and time for a self-paced 1000-m DP test (CV = 1.3%)[8] were somewhat higher than for the times of two self-paced skating tests over 1000 m (CV = 2.7%)[10] and 800 m (CV = 3.6%).[14] Reliability for $\sum O_2$-deficit, an estimate for the anaerobic capacity calculated based on the above-mentioned 1000-m and 800-m tests, showed CVs of 8.1%[10] and 9.8%.[14] Among elite cross-country skiers, the CV values for the strength variables of a two-phase ski-specific upper-body strength test ranged between 1.4% and 6.8% for a four-repetition maximum test and between 1.1% and 7.0% for a 40-repetition test.[9] To our knowledge, test–retest reliability measures for key performance indicators in XC skiing, such as $\dot{V}O_{2max}$, performance at the second lactate threshold, and aerobic time trials, have not been assessed.

Therefore, the present study aimed to investigate the test–retest reliability of XC ski-specific tests within the aerobic, sprint, and neuromuscular domains. The tests are all based on the official laboratory test protocols used for the biannual performance assessments of the Swiss XC ski

national team to (1) assess key performance indicators, (2) establish intensity zones for endurance training, and (3) measure standardized skiing performance. In addition, the minimum detectable change (MDC) of performance measures was investigated so that practitioners could differentiate between typical between-day variations and actual performance changes.

2 | MATERIALS AND METHODS

2.1 | Participants

Thirty-nine Swiss XC skiers (26 men and 13 women) volunteered to participate in this study involving two independent measurement periods, with 16 (about 40%) of the athletes completing both (Table 1). Ten athletes were classified as elite and 29 as highly trained,[19] including seven Swiss Junior or U23 national team skiers. Due to a preemptive Covid-19 quarantine upon exposure during the trials, one skier was excluded from the analysis of the first measurement period. Before participating in the study, all skiers received a health assessment and were found to be free of injury and disease. Two skiers unable to complete both 24-min DP time trials due to sickness and one skier unable to adhere to the treadmill sprint test protocol during the second trial were excluded from the analysis of the corresponding test. The Regional Ethics Committee in Berne, Switzerland, approved the study (Study ID: 2020–00925), and all skiers (or their parents if they were younger than 18 years) provided written consent to participate. Participants were free to withdraw at any time during the study without having to provide a reason for their decision.

2.2 | Study design

This test–retest analysis included two measurement periods over 12 months during the skiing off-season phase (period 1: August–October, period 2: April–July). The participants completed two identical test trials for five different performance tests, with the trials separated by 2–4 days to allow for adequate recovery following the first trial and to keep the stay at the testing facility compatible with their training schedule and hence increase adherence to the protocol (Figure 1). During the first measurement period, the skiers performed a graded exercise test (GXT) in the morning, followed by a $\dot{V}O_{2max}$ test in the afternoon. During the second period, the test protocol in the morning consisted of an upper-body strength test (UB-ST), a peak power test (PP_{Erg}), and a three-component sprint test. In the afternoon, the skiers completed a 24-min double poling time trial (24-min DP). Between test trials, athletes were instructed to perform two light-intensity exercise sessions (running, roller skiing, or cycling) and one core strength session, each lasting 60–90 min according to the athlete's training history and tolerance.

2.3 | Procedures

2.3.1 | Test preparation and anthropometrics

Participants were instructed to adopt the same precompetition preparation before both testing sessions, including dietary intake, hydration state, and caffeine consumption. The skiers were required to refrain from strenuous exercise within 24 h before the trials. To reduce the impact of the circadian rhythm, trial 1 and trial 2 were scheduled for the same time of day (±1 h). Standardized laboratory conditions were maintained and controlled throughout the study period (air temperature: 18.5 ± 1.7°C; relative humidity: 54 ± 11%; barometric pressure: 911 ± 7 mmHg). Skiers unaccustomed to the test procedures completed a separate familiarization trial in the weeks before the trial. Approximately 50% of the participating skiers had previously performed the test protocols multiple times as part of routine testing as members of

TABLE 1 Participant characteristics for measurement period one and two.

	Measurement period 1		Measurement period 2	
	Females	Males	Females	Males
n	10	17	10	17
Age (y)	21.5 ± 3.7	21.6 ± 3.6	22.1 ± 2.9	23.4 ± 4.9
Body height (cm)	168.1 ± 4.8	180.8 ± 4.8	169.6 ± 4.8	178.8 ± 3.8
Body mass (kg)	61.9 ± 8.5	73.0 ± 4.4	62.2 ± 6.5	71.9 ± 5.2
Body fat (%)	20.8 ± 4.5	11.0 ± 2.0	20.9 ± 4.6	10.9 ± 2.2
$\dot{V}O_{2max}$ ($mL \cdot kg^{-1} \cdot min^{-1}$)	58.8 ± 4.4	70.1 ± 4.5	n.a	n.a

Note: Data presented as mean ± standard deviation (SD). The $\dot{V}O_{2max}$ test data are only available for the first measurement period.

Abbreviation: $\dot{V}O_{2max}$, maximum oxygen consumption.

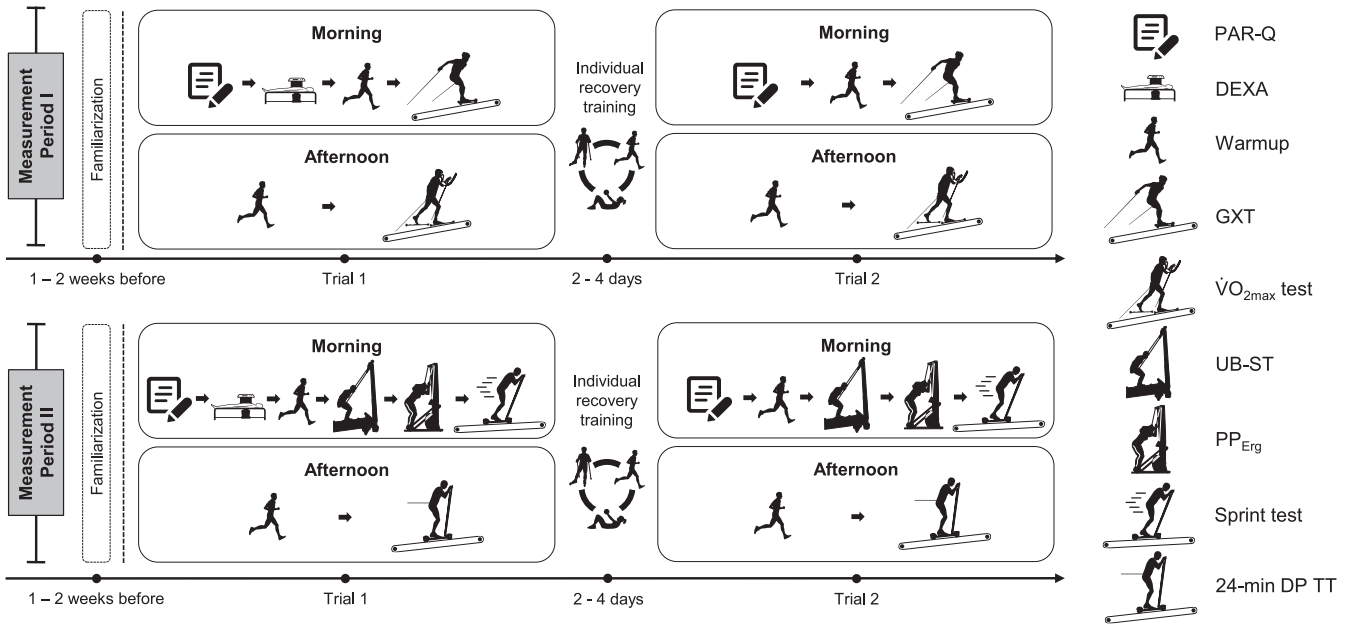

FIGURE 1 Test–retest study design involving two separate measurement periods. 24-min DP TT, 24-min double poling time trial; DXA, dual-energy X-ray absorptiometry; GXT, graded exercise test; PAR-Q, physical activity readiness questionnaire; PP_{Erg}, SkiErg peak power test; UB-ST, upper-body strength test.

the national team (measurement period 1: 13 of 27 skiers, measurement period 2: 12 of 27 skiers with experience of a specific test form). Body mass, lean body mass, and fat mass were determined during a 10-min measurement in the supine position using dual-energy X-ray absorptiometry (DXA; Lunar iDXA, GE Medical Systems, Chicago, IL, USA). Trained medical personnel conducted these examinations at the Swiss Olympic medical facility on the morning of the first trial during each measurement period. Before each test session, the skiers performed a general, standardized 15-min running warmup at 70% of the maximum heart rate, followed by a specific warmup for each test form described below. The skiers did not receive feedback on their performance (time, current velocity, or power output), except for the 24-min DP and treadmill sprint tests. The remaining stage time was provided for pacing purposes for these two tests. The same experienced test instructor conducted both trials for each test.

2.3.2 | $\dot{V}O_{2max}$ test

$\dot{V}O_{2max}$ was determined on the treadmill in the diagonal technique. Before the test, skiers completed two 2-min sub-maximum warmups at a treadmill incline of 9° with progressive intensity, separated by 3 min of active recovery at 1° and 2.80 $m \cdot s^{-1}$ for men and 2.50 $m \cdot s^{-1}$ for women. The $\dot{V}O_{2max}$ test started at an incline of 9° and a treadmill velocity of 2.40 $m \cdot s^{-1}$ for men and 1.90 $m \cdot s^{-1}$ for women, with a stepwise velocity increase every 60 s of 0.10 $m \cdot s^{-1}$ (0.15 $m \cdot s^{-1}$ for men during the first two stages) for the first 3 min, followed by an increase every 30 s until task failure. Task failure was defined as the moment when the skier could not maintain a position within 2 m of the front of the treadmill, marked by an elastic cord stretched across the treadmill. Expired air was collected at 30-s intervals via a two-way non-rebreathing valve (Hans Rudolph Inc., Shawnee, KS, USA) and plastic tubing into 100 L Douglas bags (Cranlea Human Performance Ltd, Birmingham, United Kingdom). The filling of bags started approximately 3 min before the anticipated test termination, resulting in a mean number of bags per athlete of 6.4 ± 1.4. Expiratory gas concentrations and volume were analyzed immediately after each test using a custom-built Douglas bag metabolic cart. Before a series of bags was analyzed, high-precision O_2 and CO_2 dry gas analyzers (S-3A/I and CD-3A, AEI Technologies, Inc., Bastrop, TX, USA) were calibrated using two-point calibration with precision-analyzed gas mixtures. The volume of expired air collected in the Douglas bag was determined by extracting the air through a dry gas volume meter (Hugo Sachs Elektronik GmbH, March-Hugstetten, Germany) with a vacuum pump (Cranlea Human Performance Ltd, Birmingham, United Kingdom) at a flow rate of 90 $L \cdot min^{-1}$. Corresponding fractions of expired O_2 (F_EO_2) and CO_2 (F_ECO_2) were determined simultaneously at a flow rate of 175 $mL \cdot min^{-1}$ using the gas analyzers and a flow control unit (R-1 Flow Control, AEI Technologies, Inc., Bastrop, TX, USA). $\dot{V}O_2$ and $\dot{V}_E$ were

calculated using F_EO_2, F_ECO_2, and bag volume, considering ambient O_2 and CO_2 concentrations, air temperature, relative humidity, and barometric pressure. $\dot{V}O_{2max}$ was defined as the highest oxygen uptake measured in one bag, meeting the criteria of (1) a filling time of >25 s and (2) an increase in $\dot{V}O_2$ from the previous bag of <150 mL·min^{-1}.[20] Furthermore, heart rate (HR) was measured continuously throughout the test, and the peak HR (HR_{peak}) was used for the analysis. A rating of the perceived exertion (Borg-RPE) on a 6–20 scale[21] was provided immediately after the test termination. At the same time, the test instructor collected a capillary blood sample from the skier's ear lobe to analyze the peak blood lactate concentration (BLa_{peak}).

2.3.3 | Graded exercise test (GXT)

The GXT was performed on a treadmill using skating roller skis and was preceded by a 10-min warmup on the first test stage. All skiers started at a constant velocity of 2.50 m·s^{-1} and a treadmill incline of 1°, with a 5-min stage duration and 1-min breaks between consecutive stages. The treadmill incline increased by 1° for each successive stage until it reached 8°, at which point only the treadmill velocity increased by 0.25 m·s^{-1} until task failure. Skiers chose their preferred skating sub-technique, G3 or G2, in the first trial but were asked to adopt the same skiing technique for the second trial. The same sub-technique was prescribed to avoid potential effects on the threshold determination, as the skiing techniques during the GXT can significantly impact the HR and BLa response due to differences in muscle activation patterns. HR was continuously measured and averaged over the last 60 s of each stage. During the breaks, the test instructor collected a capillary blood sample to determine the BLa, while the skiers reported their Borg-RPE. HR and BLa measured at task failure were defined as HR_{peak} and BLa_{peak}, respectively. The first and second lactate thresholds (LT_1 and LT_2) were calculated using the modified D_{max} method[22] in a custom-made Excel spreadsheet. The following formula was used to calculate the test performance:

$$\text{Test performance (stage)} = \text{Last completed stage} + \frac{\text{Time completed on stage (s)}}{300\ \text{s}} \quad (1)$$

Two experienced investigators independently evaluated the detection of LT_1 by the software algorithm, defined as a rise in BLa of >0.4 mmol·L^{-1} between consecutive stages[22] and modified to the specific workload increments in the GXT. They corrected the software's automatic detection of LT_1 by 0.5 GXT performance steps if indicated by performance, HR, or Borg-RPE relative to the corresponding maximum values of the participant and consulted a third investigator if the definition of LT_1 deviated between the two investigators. LT_2 was defined as the point on the third-order polynomial curve that produced the maximum perpendicular distance to the straight line formed by the point at LT_1 and the BLa measurement at the test termination.[22]

2.3.4 | 24-min double poling time trial (24-min DP)

Ski-specific endurance performance was measured during a 3 × 8 min self-paced treadmill time trial using the DP technique on roller skis. Before the 24-min DP, the skiers completed a 15-min standardized warmup that included 10-min DP at a 2° treadmill incline and 2.8 m·s^{-1} for men and 2.5 m·s^{-1} for women, followed by 2 min at 3° and with the individual test velocity. Each 8-min interval started with 5 min at a 3° treadmill incline, followed immediately by 3 min at 5°, with a 2-min passive recovery between intervals. The treadmill velocity was determined by (1) the pre-defined speed settings based on the mean skiing velocity during the familiarization trial and (2) the position of the skier on the treadmill, resulting in a change in velocity using a position-based, linear speed gradient. The skier controlled the position-based velocity via a wire displacement sensor (Wiege-Data, Leipzig, Germany) originating at the rear end of the treadmill and attached to a safety hip belt worn by the skier. The test performance was determined by the accumulated total distance covered over the 24-min timeframe. HR was continuously measured and averaged over the final 60 s of each 3° and 5° segment in the three intervals.

2.3.5 | Treadmill sprint test

Ski-specific sprint capacity was measured using a three-component treadmill test on roller skis. Before the test, the skiers completed a standardized 15-min warmup that included (1) a 30-s DP effort at a 2° treadmill incline starting at 5.50 m·s^{-1} for men and 4.45 m·s^{-1} for women, with progressively increasing speed; (2) a 30-s skating effort at a treadmill inclination of 3° and a velocity of 6.50 m·s^{-1} for men and 5.00 m·s^{-1} for women; and (3) a sub-maximum familiarization trial of the first test component ($Sprint_{DP1}$), with all three warmup parts separated by ~3-min rest periods. The first test component ($Sprint_{DP1}$) measured DP peak velocity ($Sprint_{DP1}$

V_{peak}) using a protocol previously introduced by Carlsson et al.[4] Within a 10-s period at a fixed incline of 2°, the treadmill velocity accelerated to 5.50 m·s^{-1} for men and 4.45 m·s^{-1} for women, after which a stepwise velocity increase of 0.28 m·s^{-1} was applied every 4 s until task failure. The test terminated when the skier could not maintain the roller ski front wheels ahead of a virtual line placed 2 m from the front of the treadmill, indicated by a marker on the side of the treadmill. The second test component ($Sprint_{1\text{-}min}$) was performed in the skating technique and followed immediately after a 60-s passive break, during which the poles were switched to a length appropriate for skating. At a constant treadmill incline of 3° and after an initial 20-s set at a velocity of 4.50 m·s^{-1} for men and 3.50 m·s^{-1} for women, the treadmill velocity increased to 6.50 m·s^{-1} for men and 5.00 m·s^{-1} for women. Following the initiation of this fixed velocity, the skiers controlled the treadmill speed via their position on the treadmill using the wire displacement sensor. At a self-selected pace using the skating G3 subtechnique, the skiers attempted to maintain their maximum mean velocity over 60 s ($Sprint_{1\text{-}min}$ V_{avg}). They repeated the first test component after another 60-s passive recovery to collect a capillary blood sample and Borg-RPE and to change back to classic-length poles. This third test component was termed $Sprint_{DP2}$ to establish the peak velocity in a fatigued state ($Sprint_{DP2}$ V_{peak}). The DP peak velocity for $Sprint_{DP1}$ and $Sprint_{DP2}$ was determined as the treadmill speed recorded during the last 4-s stage, during which at least half was completed. For $Sprint_{1\text{-}min}$, the mean velocity during the 60 s was the performance indicator. HR was continuously measured, and HR_{peak} was extracted for $Sprint_{1\text{-}min}$. For safety reasons, the skiers wore a chest harness connected to the emergency brake above the treadmill.

2.3.6 | Upper-body strength test (UB-ST)

Ski-specific explosive upper-body strength was determined using a custom-made device (in-house validation, unpublished), with which the skiers performed single, maximum pulls in the DP motion with light (LIT), medium (MIT), and heavy (HIT) resistance. The skiers were positioned on a ramp wearing roller skis (Marwe Skating 610 A, US7, Marwe OY, Hyvinkää, Finland) at an incline of 6°, 8°, and 10° for LIT, MIT, and HIT, respectively. They pulled on a pulley system in front of them while wearing XC ski straps attached to two parallel ropes. The pulley system slid downward during the poling motion, simulating the dynamic pole force vector observed during DP on skis. Between the pole straps and the rope, two one-component force cells (KM26z, ME-Messsysteme, Henningsdorf, Germany) measured the left and right stroke forces at a sample rate of 250 Hz. A linear position transducer (SX800-2500-1R-KA, WayCon Positionsmesstechnik GmbH, Brühl, Germany) measured vertical displacement at 250 Hz on the pulley sled. The starting position of the sled carrying the load was determined based on body height. From a tucked position, imitating the end of the poling recovery phase on skis, the skiers quickly repositioned themselves in a forward-leaning upright stance, after which they performed a maximum downward pull. They completed five maximum repetitions for each of the three incremental loads (LIT, MIT, and HIT), separated by 30 s of passive recovery. Loads applied to the pulley sled were individually determined according to body mass and ranged from 3 to 50 kg for the three load categories based on pilot testing. Mean force (F_{mean}) and peak power (P_{peak}) were calculated using a custom-made Python script for each load condition and were reported as the mean of three trials; the lowest and highest attempts were dropped.

2.3.7 | SkiErg peak power test (PP_{Erg})

DP peak power was measured on a modified Concept2 ski ergometer (SkiErg; Concept2, Morrisville, VT, USA) with the skier fixed to XC ski bindings on adjustable rails in front of the ergometer. For the test, the skiers performed two 20-s trials separated by a 3-min recovery phase (1-min passive, 1-min active, and 1-min passive). The skiers were instructed to perform six introductory strokes, followed by all-out strokes, according to the peak power protocols with similar equipment.[23] To scale for body mass, the drag factor settings of the ergometer were adjusted via a flywheel damper to 110% and 130% of the body mass for women and men, respectively. An elastic resistance band (length: 208 cm, resistance: 25–30 kg) was placed around the skier's hips. A backward pull was applied at the skier's preferred tension to facilitate the forward-leaning motion observed during on-snow DP. The position of the bindings, the damper setting, and the tension of the elastic band were determined during the familiarization trial and replicated for all subsequent trials. Before the test, the skiers performed a 5-min warmup at 1.0–1.5 W·kg^{-1} interspersed with two 6-s sprints. Then, a sub-maximum test trial was conducted at 90% maximum effort, followed by another 2-min recovery before the two maximum effort trials. Ergometer data were recorded stroke by stroke using the Concept2 mobile application (ErgData, v.1.95, Concept2, Morrisville, VT, USA) and then exported to a comma-separated values (CSV) file for analysis. Peak

power was the highest moving average over five consecutive strokes recorded during both trials.

2.3.8 | Materials and equipment

All treadmill tests were conducted on a large motorized treadmill (3 × 4.5 m, Poma, Porschendorf, Germany) using either classic roller skis (C2, Swix, Lillehammer, Norway) for the $\dot{V}O_{2max}$ test or skating roller skis (Marwe Skating 610 A, wheel type US6/7, Marwe OY, Hyvinkää, Finland) for the GXT, 24-min DP, treadmill sprint test, and UB-ST. All skiers used XC ski poles (Triac 3, Swix, Lillehammer, Norway) equipped with custom-made pole tips for treadmill use. Roller skis were warmed up for a minimum of 10 min before each test.[18] The average power output for the 24-min DP and $Sprint_{1\text{-min}}$ and the peak power output for $Sprint_{DP1}$ and $Sprint_{DP2}$ were calculated as the sum of power against gravity and rolling friction:

$$P_t = P_g + P_f + P_w \quad (2)$$

$$P_g = m \times g \times v \times \sin(\alpha) \quad (3)$$

$$P_f = m \times g \times v \times \cos(\alpha) \times \mu \quad (4)$$

$$P_w = F_w \times v \quad (5)$$

where P_t is the total power output, P_g is the work against gravity, P_f is the work against rolling friction, P_w is the work against the wire displacement sensor, m is the mass of the skier including equipment, g is the gravitational constant, v is the treadmill velocity, α is the inclination of the treadmill, μ the coefficient of friction, and F_w the resistive force of the displacement wire. μ for all roller ski models was determined before, during, and after the two test periods weekly via the tow test on the treadmill by the same investigator as previously described.[24] With a constant treadmill velocity of 5 m·s^{-1} and an incline of 0°, the investigator (70 kg body mass) rolled passively for 15 min while connected to a strain gauge (KD80s, ME-Messsysteme GmbH, Hennigsdorf, Germany). Tow force was measured continuously at 1 Hz and averaged over the last 2 min. The μ values (mean ± SD) for the roller skis were 0.02388 ± 0.00075 for the Marwe US6, 0.02631 ± 0.00035 for the Marwe US7, and 0.02046 ± 0.00002 for the Swix Classic C2. F_w was measured with the same strain gauge, with a resistance force of 4.5 N. All HR measures were monitored using a chest belt and analyzed with Firstbeat Sports (Firstbeat Technologies Oy, Jyväskylä, Finland). BLa was determined via capillary blood collected from a 20-µL sample taken from the skier's ear lobe, hemolyzed in a pre-filled micro-test tube, and analyzed in a lactate analyzer (Biosen C-Line, EKF Diagnostics, Barleben, Germany).

2.4 | Statistical analysis

Test and retest data are reported as mean ± SD. The normal distribution of all data was assessed visually using quantile-quantile (Q-Q) plots and Shapiro–Wilk test statistics. The equality of variance between the trial 1 and trial 2 samples was confirmed using Levene's test, with all data demonstrating $p > 0.05$. Paired-sampled *t*-tests, or Wilcoxon signed-rank tests in the case of non-parametric data, were performed to determine differences in outcome measures between the two trials to identify any potential systematic familiarization bias.[12] Cohen's d effect sizes (ESs) with Hedges correction for paired samples were calculated to evaluate the practical significance of these differences. The effect size interpretation proposed by Cohen[25] was applied, with 0.2 = small effect, 0.5 = moderate effect, and 0.8 = large effect. Relative reliability was calculated for all variables using intraclass correlation coefficient (ICC) estimates and their 95% confidence intervals (CIs) based on a two-way mixed single measure ($ICC_{3,1}$), according to Hopkins.[26] Classification of ICCs was conducted according to Koo and Li[27] with the correlation estimates <0.5 = poor, 0.5–0.75 = moderate, 0.75–0.9 = good, and >0.9 = excellent. Absolute reliability indicators for the outcome measures were calculated as the standard error of measurement (SEM) in absolute values, according to Weir,[28] where SDd represents the standard deviation of the difference scores:

$$\text{SEM} = \frac{\text{SDd}}{\sqrt{2}} \quad (6)$$

The CV of the within-subject variation was derived from the standard deviation of the log-transformed differences (SDd_{log}) and complemented with a 95% CI, as outlined by Hopkins[26]:

$$\text{CV} = \frac{\text{SDd}_{\log}}{\sqrt{2}} \times 100 \quad (7)$$

The CV as a reliability measure is specific to the variable being tested, and there are no generally accepted thresholds for the qualitative assessment of CV values.[29] Based on previously proposed CV values ≤5% for fitness testing,[30] together with the performance caliber and the expected performance variation of the skiers, we defined CV scores as follows: >5.0 = poor, 5.0–2.5 = moderate, 2.5–1.0 = good, and <1.0 = excellent. The MDC at the 80%

CI was calculated in absolute (8) and relative terms (9), where $\overline{X}$ represents the mean for all observations from trials 1 and 2:

$$\mathrm{MDC}_{80} = \mathrm{SEM} \times 1.28 \times \sqrt{2} \quad (8)$$

$$\text{relative MDC}_{80} = \frac{\mathrm{MDC}_{80}}{\overline{X}} \times 100 \quad (9)$$

The MDC can be used to practically interpret the change required in measurements to have 80% certainty that a real change has occurred.[28] We opted for the MDC_{80} instead of the MDC with 95% certainty, in agreement with Ettema et al.[31] and the proposal to consider less conservative decision limits with athletes to avoid unrealistic changes in performance.[32] Consequently, any change that exceeds the MDC would be meaningful. All reliability analyses were performed using R v.4.0.2.[33] The statistical significance threshold was set at $p < 0.05$.

3 | RESULTS

3.1 | Aerobic tests

The absolute $\dot{V}O_{2max}$ demonstrated relative and absolute test–retest reliability of ICC = 0.99 and CV = 1.4%, respectively, and an MDC of 112 mL·min^{-1} (Table 2 and Figure 2A,B). The maximum test performance and the performance at LT_2 during the GXT revealed higher relative (ICC >0.95) and absolute (CV = 2–3%) reliability measures compared to the performance at LT_1 (ICC = 0.80, CV = 8.7%) (Figure 2C,D). Compared to the LT_1 HR, LT_2 HR, and HR_{peak} during the GXT (ICC = 0.86–0.94; CV = 1.0–2.7%), the LT_1 BLa, LT_2 BLa, and BLa_{peak} showed lower reliability (ICC = 0.58–0.82; CV = 7.7–14.4%). The skiers' performance during the 24-min DP showed relative and absolute reliability estimates of ICC = 0.99, CV = 1.0%, and SEM = 49 m (Figure 2E,F). The relative MDC values ranged from 1.9% to 6.5% for the test performance of the $\dot{V}O_{2max}$ test, GXT, and 24-min DP. The GXT and 24-min DP test performance demonstrated significant differences between trials 1 and 2 (both $p < 0.05$). Significant differences between trials 1 and 2 were also observed for the LT_1 BLa, LT_2 BLa, and HR_{peak} during the GXT (all $p < 0.05$). Analyses of the effect sizes revealed trivial to moderate differences between trials 1 and 2 for the $\dot{V}O_{2max}$ test time to task failure and the GXT test performance (ES = 0.36–0.75) and large differences for the 24-min DP performance (ES = 1.26).

3.2 | Sprint and neuromuscular tests

Test–retest reliability estimates for the treadmill sprint performance outcomes, Sprint_{DP1} V_{peak}, $\mathrm{Sprint}_{1\text{-}min}$ V_{avg}, and Sprint_{DP2} V_{peak}, demonstrated relative and absolute reliability of ICC >0.96 and CV <2.3%, respectively, and relative MDC values ranging from 2.4% to 4.4% (Table 3). Sprint_{DP2} V_{peak} was significantly higher during trial 2 compared to trial 1 ($p = 0.037$, ES = 0.44). UB-ST F_{mean} and P_{peak} across LIT, MIT, and HIT loads exhibited relative reliability measures of ICC >0.90 and absolute reliability of CV = 4.4–7.8%. SkiErg P_{peak} displayed absolute and relative reliability of ICC = 0.99 and CV = 2.4%, respectively. A comprehensive list of all outcome measures can be found in Appendix S1.

4 | DISCUSSION

In the present study, we investigated the test–retest reliability of a comprehensive XC ski-specific performance test battery that included aerobic, sprint, and neuromuscular performance tests in highly trained XC skiers. We demonstrated (1) good-to-excellent reliability for $\dot{V}O_{2max}$ and 24-min DP performance, (2) poor-to-moderate reliability for performance at the first (LT_1 stage), and moderate-to-excellent reliability for performance at the second (LT_2 stage) lactate threshold during the GXT, (3) good-to-excellent reliability for peak-velocity DP sprint and 1-min self-paced skating sprint performance on the treadmill, (4) excellent relative but poor absolute reliability for upper-body explosive strength and power, and (5) good-to-excellent reliability of ergometer-derived peak power based on the reliability criteria. Using treadmill-based protocols and ski ergometers, highly trained XC skiers may assess ski-specific key performance indicators in the endurance, sprint, and power domains with reasonable reliability. Whether the presented test forms are valid regarding on-snow performance remains to be determined in subsequent studies.

4.1 | Aerobic tests

4.1.1 | $\dot{V}O_{2max}$ test

$\dot{V}O_{2max}$ revealed excellent relative (ICC = 0.99) and good absolute (CV = 1.4%) reliability during the treadmill diagonal skiing to task failure. To our knowledge, test–retest reliability for a treadmill-based assessment of $\dot{V}O_{2max}$ on roller skis has not been previously reported. Although the Douglas bag method has demonstrated

TABLE 2 Test–retest reliability statistics for the aerobic exercise tests and physiological characteristics in highly trained male and female cross-country skiers.

Outcome measure	Mean ± SD Trial 1	Mean ± SD Trial 2	*p*-Value	Effect size	ICC [95% CI]	SEM [95% CI]	CV% [95% CI]	MDC_{80}	$\%MDC_{80}$
$\dot{V}O_{2max}$ test									
TTTF (s)	350 ± 56	356 ± 57	0.067	0.36	0.95 [0.90–0.97]	13 [10–17]	3.7 [2.9–5.0]	23	6.5
$\dot{V}O_{2max}$ ($mL \cdot min^{-1}$)	4541 ± 798	4565 ± 807	0.307	0.26	0.99 [0.99–1.00]	62 [49–85]	1.4 [1.1–1.9]	112	2.5
$\dot{V}O_{2max}$ ($mL \cdot kg^{-1} \cdot min^{-1}$)	65.8 ± 7.0	66.1 ± 7.3	0.145	0.28	0.99 [0.97–0.99]	0.9 [0.7–1.2]	1.3 [1.0–1.8]	1.5	2.3
$\dot{V}_E$ ($L \cdot min^{-1}$)	172 ± 32	173 ± 32	0.883	0.03	0.97 [0.95–0.98]	5 [4–8]	3.0 [2.4–4.1]	10	5.8
RER	1.13 ± 0.04	1.12 ± 0.03	0.138	0.29	0.67 [0.44–0.81]	0.02 [0.02–0.03]	1.8 [1.4–2.5]	0.04	3.3
HR_{peak} (bpm)	193 ± 8	192 ± 8	0.125	0.30	0.96 [0.92–0.98]	2 [1–2]	0.8 [0.6–1.1]	3	1.5
BLa ($mmol \cdot L^{-1}$)	7.18 ± 1.45	7.14 ± 1.41	0.882	0.03	0.70 [0.49–0.83]	0.80 [0.63–1.10]	11.8 [9.3–16.2]	1.45	20.2
Graded exercise test									
Test performance (1–12)	7.80 ± 1.08	8.01 ± 1.18	<0.001	0.75	0.97 [0.95–0.99]	0.19 [0.15–0.26]	2.3 [1.8–3.1]	0.34	4.3
LT_1 stage (1–12)	2.85 ± 0.50	2.98 ± 0.70	0.097	0.33	0.80 [0.65–0.89]	0.27 [0.21–0.37]	8.7 [6.8–11.9]	0.49	16.8
LT_1 BLa ($mmol \cdot L^{-1}$)	1.43 ± 0.27	1.64 ± 0.37	0.002	0.66	0.58 [0.32–0.76]	0.21 [0.17–0.29]	14.4 [11.4–19.8]	0.38	24.9
LT_1 HR (bpm)	144 ± 10	144 ± 10	0.713	0.07	0.86 [0.74–0.92]	4 [3–5]	2.7 [2.1–3.6]	7	4.8
LT_2 stage (1–12)	5.76 ± 0.82	5.94 ± 0.95	0.003	0.62	0.95 [0.91–0.97]	0.20 [0.16–0.27]	3.1 [2.5–4.3]	0.36	6.1
LT_2 BLa ($mmol \cdot L^{-1}$)	3.64 ± 0.61	3.86 ± 0.73	0.014	0.49	0.80 [0.64–0.89]	0.30 [0.24–0.42]	7.7 [6.1–10.6]	0.55	14.7
LT_2 HR (bpm)	179 ± 7	179 ± 8	0.817	0.04	0.93 [0.87–0.96]	2 [2–3]	1.2 [0.9–1.6]	4	2.1
HR_{peak} (bpm)	191 ± 8	192 ± 7	0.021	0.46	0.94 [0.88–0.97]	2 [1–3]	1.0 [0.8–1.4]	3	1.8
BLa_{peak} ($mmol \cdot L^{-1}$)	10.30 ± 1.91	10.60 ± 2.22	0.162	0.27	0.82 [0.69–0.90]	0.87 [0.68–1.19]	7.7 [6.1–10.6]	1.57	15.1
RPE (6–20)	19.7 ± 0.7	20.0 ± 0.2	0.048	0.43	0.45 [0.16–0.67]	0.4 [0.3–0.5]	1.9 [1.5–2.6]	0.7	3.3
24-min DP time trial									
Distance (m)	4687 ± 486	4777 ± 497	<0.001	1.26	0.99 [0.98–0.99]	49 [38–68]	1.0 [0.8–1.4]	89	1.9
P_{avg} (W)	218 ± 41	223 ± 41	<0.001	1.26	1.00 [0.99–1.00]	3 [2–4]	1.2 [0.9–1.7]	5	2.3
${}_{rel}P_{avg}$ ($W \cdot kg^{-1}$)	3.20 ± 0.36	3.26 ± 0.35	<0.001	1.08	0.99 [0.98–0.99]	0.04 [0.03–0.05]	1.2 [1.0–1.7]	0.07	2.2

Note: Data are presented as mean ± standard deviation (SD). *p*-value from mean difference test statistic between trial 1 and trial 2. Effect size between group means as Cohen's d. $N = 27$ for $\dot{V}O_{2max}$ test and graded exercise test; $N = 25$ for 24-min DP time trial.

Abbreviations: BLa, blood lactate concentration; CI, confidence interval; CV, coefficient of variation; DP, double poling; HR_{peak}, peak heart rate; ICC, intraclass correlation coefficient; LT_1, first lactate threshold; LT_2, second lactate threshold; MDC_{80}, minimal detectable change at 80% confidence interval; P_{avg}, average power output; ${}_{rel}P_{avg}$, relative average power output; RER, respiratory exchange ratio; RPE, rating of perceived exertion; SEM, standard error of measurement; TTTF, time to task failure; $\dot{V}_E$, ventilation; $\dot{V}O_{2max}$, maximum oxygen uptake.

high reliability for the measurement of $\dot{V}O_{2peak}$ during ski-specific exercise (CV = 1.9–2.4%; r = 0.99–1.00), the exercise protocols were conducted on ski ergometers with <10 participants in these two studies.[15,17] At the same time, the reliability of $\dot{V}O_{2peak}$ assessed during self-paced treadmill time trials in the skating technique demonstrated slightly lower values (CV = 2.3–3.6%) compared to our findings.[10,14] When comparing the reliability of $\dot{V}O_{2peak}$ and $\dot{V}O_{2max}$ assessed during roller skiing to other exercise modes, the reliability appears similar in running (CV = 2.8%).[34] Given the high reliability of $\dot{V}O_{2max}$ in the present study, we suggest using the Douglas bag method during a ski-specific, incremental treadmill protocol on roller skis to measure $\dot{V}O_{2max}$ in elite XC skiers and biathletes. First, $\dot{V}O_{2max}$ has been consistently shown to be one of the most important performance indicators in XC skiing[3–7] and should be assessed preferably during the actual skiing motion.[18] In contrast to diagonal skiing on the treadmill, the DP ergometer allows only limited on-snow transfer and involves less total muscle mass. The test mode for assessing $\dot{V}O_{2max}$ should ideally incorporate the maximum activity of upper-body and leg muscle mass, such as observed during diagonal skiing,[35] to capture an athlete's true aerobic potential. Second, the smallest worthwhile enhancement in performance, defined as 0.3 times the within-athlete variability,[36] lies between 0.3% and 0.7% for elite XC skiers.[11] To detect such small practical important changes, the measurement error in time trials and incremental tests performed in the laboratory should not exceed the on-snow performance variability.[11] Whether $\dot{V}O_{2max}$ assessed during roller skiing and running is equivalent for XC skiers needs further investigation, as many elite XC skiers determine their $\dot{V}O_{2max}$ through running protocols.

4.1.2 | Graded exercise test

Performance at LT_1 and LT_2 during the GXT demonstrated poor-to-moderate absolute reliability (CV = 8.7% and 3.1%, respectively). The significant differences in the GXT test performance and LT_2 between the two trials further suggest that a systematic bias was present (i.e., learning effect, motivation, or fatigue). However, the effect size for the difference was moderate (ES = 0.62–0.75), and the higher mean performance in trial 2 was influenced strongly by a few individuals performing better during the second trial. Previous studies reporting on the reliability of performance at LT_1 (defined here as the workload associated with a lactate concentration of 0.5 mmol L^{-1} above resting level) in trained individuals have shown lower CV values of 3.7% for cycling[37] and 2.0% for running,[34] as compared to the present study. Reports on the D_{max} method to identify LT_2 in other exercise modes showed high reliability (CV = 2.1% and ICC = 0.94) for running[34] but lower and more varying reliability values (CV = 3.8–10.3% and ICC = 0.57–0.90) for cycling[37,38] compared to the present results. Reliability estimates for the modified D_{max} methods to determine LT_2 have not been reported to our knowledge and, therefore, might differ from the reliability of the D_{max} method. Our reliability results for LT_1 and LT_2 may be attributed to the stage-based discontinuous exercise protocol and the inherent variability of BLa as the parameter defining both thresholds. First, upon completing a stage during the protocol, the skiers could complete a considerable amount of the subsequent exercise stage due to the 60-s recovery periods. Although the skiers were blinded to the time elapsed during the stage, they could easily count the number of stages. Our inability to effectively blind the skiers might have led to considerable differences between the two trials for LT_2, as the determination of LT_2 using the modified D_{max} method considers the lactate point at the test termination. Second, the indirect deduction of the performance variable based on an internal load parameter (i.e., BLa) contains considerable variability. Although the pre-test BLa values did not differ significantly between trials 1 and 2 (1.30 mmol·L^{-1} and 1.37 mmol·L^{-1}, respectively, p = 0.363), the BLa at LT_1 and LT_2 were higher in the skiers during trial 2. To improve the reliability of BLa measures at LT_1 and LT_2, it could be argued to incorporate the delta BLa (e.g., the difference between BLa at LT1 and BLa at resting level) instead of using the absolute BLa value. However, the performance and HR at LT_1 were not significantly affected by the elevated BLa baseline in trial 2 (2.85 and 2.95 for stage at LT_1 and 144 bpm in both trials for HR at LT_1, both $p > 0.05$). The primary purpose of the GXT for XC skiers is to estimate the HR zones corresponding to the different exercise intensities for endurance training based on the increase in the BLa. The data indicated moderate-to-excellent reliability measures for HR at LT_1 and LT_2 (both ICC

FIGURE 2 Scatterplots and Bland–Altman plots showing the test–retest reliability for $\dot{V}O_{2max}$ (A and B), GXT LT_2, (C and D), and 24-min DP TT distance (E and F) in female (triangles) and male (circles) skiers. The dashed line in the scatterplots indicates the line of identity. The horizontal line in the Bland–Altman plot indicates the mean difference between trial 1 and trial 2, whereas the dashed lines represent the 95% limits of agreement.

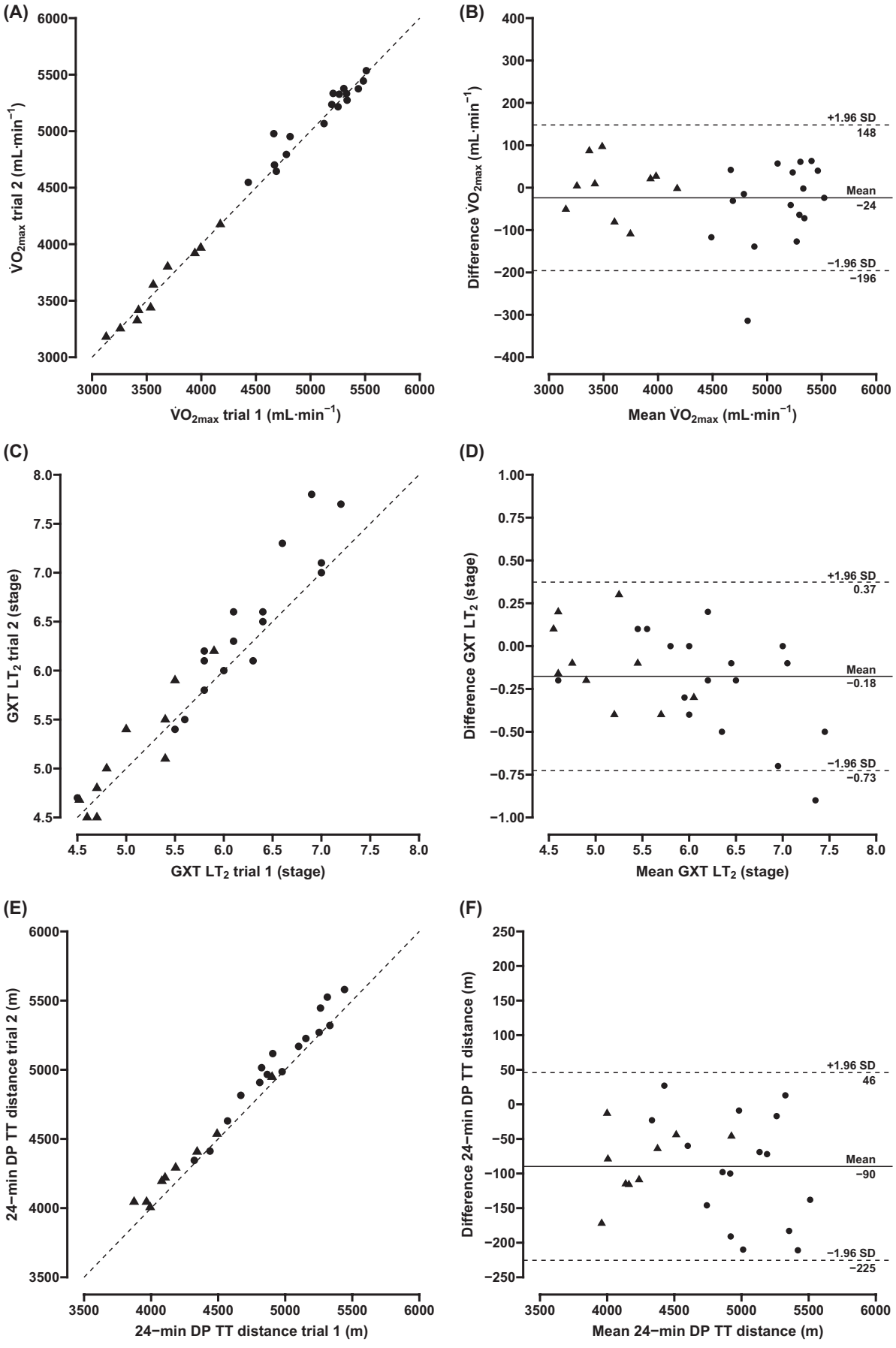

TABLE 3 Test–retest reliability statistics for the sprint and neuromuscular tests in highly trained male and female cross-country skiers.

Outcome measure	Mean ± SD Trial 1	Mean ± SD Trial 2	*p*-Value	Effect size	ICC [95% CI]	SEM [95% CI]	CV% [95% CI]	MDC_{80}	$\%MDC_{80}$
Treadmill sprint test									
$Sprint_{DP1}$ V_{peak} ($m \cdot s^{-1}$)	7.22 ± 0.74	7.27 ± 0.80	0.138	0.23	0.96 [0.92–0.98]	0.16 [0.12–0.21]	2.1 [1.7–2.9]	0.28	3.9
$Sprint_{DP1}$ P_{peak} (W)	350 ± 60	354 ± 65	0.054	0.38	0.98 [0.96–0.99]	8 [7–12]	2.3 [1.8–3.1]	15	4.4
$Sprint_{1\text{-}min}$ V_{avg} ($m \cdot s^{-1}$)	5.98 ± 0.65	5.96 ± 0.62	0.903	0.00	0.99 [0.97–0.99]	0.08 [0.06–0.11]	1.3 [1.0–1.8]	0.14	2.4
$Sprint_{1\text{-}min}$ P_{avg} (W)	360 ± 64	361 ± 64	0.223	0.24	0.99 [0.99–1.00]	5 [4–7]	1.5 [1.1–2.0]	10	2.7
$Sprint_{1\text{-}min}$ HR_{peak} (bpm)	189 ± 6	187 ± 6	<0.001	0.93	0.95 [0.91–0.97]	1 [1–2]	0.7 [0.5–1.0]	2	1.3
$Sprint_{1\text{-}min}$ BLa ($mmol \cdot L^{-1}$)	8.23 ± 1.48	7.62 ± 1.34	0.008	0.55	0.80 [0.64–0.89]	0.64 [0.50–0.89]	8.3 [6.5–11.5]	1.16	14.7
$Sprint_{DP2}$ V_{peak} ($m \cdot s^{-1}$)	6.62 ± 0.64	6.72 ± 0.64	0.037	0.44	0.96 [0.92–0.98]	0.14 [0.11–0.19]	2.0 [1.6–2.8]	0.24	3.7
$Sprint_{DP2}$ P_{peak} (W)	321 ± 54	327 ± 55	0.008	0.55	0.98 [0.97–0.99]	7 [5–9]	2.1 [1.7–2.9]	12	3.8
Upper-body strength test									
LIT F_{mean} (N)	333 ± 58	325 ± 58	0.121	0.35	0.91 [0.84–0.95]	17 [13–23]	4.9 [3.8–6.7]	31	9.3
LIT P_{peak} (W)	837 ± 238	822 ± 224	0.285	0.25	0.97 [0.94–0.98]	41 [32–56]	4.9 [3.9–6.8]	74	9.0
MIT F_{mean} (N)	408 ± 70	404 ± 62	0.397	0.16	0.92 [0.86–0.96]	18 [14–25]	4.4 [3.4–6]	33	8.1
MIT P_{peak} (W)	796 ± 234	759 ± 222	0.047	0.42	0.93 [0.87–0.96]	61 [48–83]	7.8 [6.2–10.7]	110	14.2
HIT F_{mean} (N)	504 ± 80	490 ± 73	0.051	0.38	0.90 [0.82–0.95]	24 [19 – 33]	4.6 [3.6–6.3]	44	8.8
HIT P_{peak} (W)	721 ± 229	699 ± 224	0.143	0.28	0.95 [0.90–0.97]	52 [41–72]	6.6 [5.2–9.1]	95	13.4
SkiErg peak power test									
SkiErg P_{peak} (W)	471 ± 145	471 ± 147	0.876	0.01	0.99 [0.99–1.00]	12 [10–17]	2.4 [1.9–3.3]	22	4.7
SkiErg $_{rel}P_{peak}$ ($W \cdot kg^{-1}$)	6.79 ± 1.64	6.77 ± 1.63	0.381	0.11	0.99 [0.98–0.99]	0.18 [0.14–0.24]	2.4 [1.9–3.3]	0.32	4.7

Note: Data presented as mean ± standard deviation (SD). *p*-value from mean difference test statistic between trial 1 and trial 2. Effect size between group means as Cohen's d. $N = 27$ for $Sprint_{DP1}$, upper-body strength test, and SkiErg peak power Test; $N = 26$ for $Sprint_{1\text{-}min}$ and $Sprint_{DP2}$.

Abbreviations: BLa, blood lactate concentration; CI, confidence interval; CV, coefficient of variation; F_{mean}, mean force; HIT, high load; HR_{peak}, peak heart rate; ICC, intraclass correlation coefficient; LIT, light load; MDC_{80}, minimal detectable change at 80% confidence interval; MIT, medium load; P_{peak}, peak power; $_{rel}P_{peak}$, relative peak power; SEM, standard error of measurement; SkiErg, ski ergometer double poling; $Sprint_{1\text{-}min}$, 1-min self-paced skating sprint test; $Sprint_{DP1}$, double poling sprint test 1; $Sprint_{DP2}$, double poling sprint test 2; V_{avg}, average velocity; V_{peak}, peak velocity.

>0.85; CV <3.0%). Although HR monitoring does have limitations at high intensities, this study confirms that it is a widely accepted, non-invasive, and cost-effective tool for athletes and coaches to prescribe and control the training intensity in XC skiing for continuous work in the moderate and heavy domains. Measures for increasing the reliability of the threshold assessment include controlling the exercise protocol (stage and break duration, continuous vs. discontinuous), ensuring proper test familiarization, implementing thorough quality control of the measurement equipment, and understanding the methodological characteristics of the threshold definition employed.

4.1.3 | 24-min double poling time trial

The 24-min DP distance showed excellent relative and absolute reliability (ICC = 0.99; CV = 1.0%) and is similar to the 24-min DP P_{avg} (ICC = 1.00; CV = 1.2%), suggesting that the use of both measures is appropriate depending on preferences or the study setting. This study is the first to report reliability data for an aerobic-based, XC ski-specific time trial on the treadmill. The high reliability is in alignment with the time trial performance of comparable duration (about 20–25 min) in other sports, such as for the 20-km time trial time (CV = 1.1%) in cycling[39] and the 5-km time trial time (CV = 1.7%) in running.[40] Potential reasons for the high reliability of the 24-min DP include the similarities to the training habits and the repetitive, sub-maximum movement pattern. Competitive XC skiers are used to performing aerobic interval-based training of various durations, with similar demands compared to the 3 × 8 min during the 24-min DP. At the same time, the repetitive flexion-extension motion in DP is a reproducible sub-technique. DP is comparable to the movement pattern in rowing, where high test–retest reliability has been observed during aerobic test forms, such as the 2000-m time trial.[41] Despite the familiarization trial and the extended experience with the 24-min DP test form in the case of the current and former national team skiers, some participants showed significant improvements in performance between trials similar to the GXT (Figure 2F). The explanation for this result could be twofold. First, the skiers controlled their skiing velocity during the test via their position on the treadmill belt; thus, actual performance blinding was not feasible in the present study. Second, proper pacing is crucial during this test form and is often unreliable with inexperienced skiers based on the experience of the test instructor. With additional trials, skiers tend to perform at a more even pace in the later trials, which agrees with Schabort et al.'s[41] findings for the 2000-m time trial on the rowing ergometer. Thus, even pacing improves time trial performance, as better skiers adopt a more even pacing strategy during XC ski racing.[42] Before evaluating performance changes in the context of interventions, coaches and scientists utilizing the 24-min DP test should undertake at least two familiarization trials to create a stable performance baseline.

4.2 | Sprint and neuromuscular tests

4.2.1 | Treadmill sprint test

In the sprint and neuromuscular test domains, the treadmill sprint performance measures $Sprint_{DP1}$ V_{peak}, $Sprint_{1\text{-}min}$ V_{avg}, and $Sprint_{DP2}$ V_{peak} showed moderate-to-excellent reliability (ICC = 0.96–0.99; CV = 1.3–2.1%). The high reliability in the sprint test protocols is consistent with the findings by Stöggl et al.,[8] who used a comparable time-to-task-failure test protocol similar to $Sprint_{DP1}$ in the present study and established the reliability for maximum DP velocity at CV = 1.7% and $r = 0.93$. As anaerobic capacity is a critical performance indicator in sprint skiing,[6] the self-paced $Sprint_{1\text{-}min}$ might complement short sprint-type protocols lasting <30 s (e.g., $Sprint_{DP1}$) and the longer, aerobic-based protocols (e.g., the $\dot{V}O_{2max}$ test and the GXT). Stöggl et al.'s[8] comparable but somewhat longer 1000-m treadmill sprint protocol (ski time = 166 ± 22.0 s) demonstrated the same absolute reliability found in the present study, with a CV of 1.3%. The skating 1000-m and 800-m sprint protocols by Losnegard et al.,[10,14] conducted at a 6° treadmill incline and, therefore, of longer duration (~259 s and 207 s, respectively), demonstrated slightly lower test–retest reliability for skiing time, with a CV of 2.7% and 3.6%, respectively. Reliability for $\sum O_2$-deficit, representing the anaerobic capacity and measured during the skating trials mentioned above, was considerably lower (CV = 8.1% and 9.8%, respectively) compared to the $Sprint_{1\text{-}min}$ V_{avg} of the present study. Despite the notion that sprint treadmill protocols put the athlete at a certain risk due to the high treadmill belt velocities, the present protocol appeared safe and reliable, given appropriate familiarization and safety precautions (e.g., use of a safety harness and controlled quick treadmill belt deceleration).

4.2.2 | Upper-body strength test

The reliability of UB-ST F_{mean} and P_{peak} (CV = 4.4–7.8%) in the present study was lower than that of the four-repetition maximum test on a rollerboard.[9] In the study

above, Stöggl et al.[9] found that the peak force and mean power during the active pulling phase had CVs of 4.0% and 2.5%, respectively. Potential explanations for the higher reliability in that study compared to the present findings include (1) the less complex technical demands of using the rollerboard (only the upper body contributes substantially during the pulling phase), (2) the repetitive character of the protocol (four repetitions vs. one), (3) the more difficult standardization process in the present study (sled height, ramp incline, and roller skis), and (4) differences in the signal processing thresholds. Data from a comparable one-repetition peak power test on a rowing ergometer comprising a flexion-extension rowing stroke exhibited slightly better test–retest reliability, with CV = 4.9% and ICC = 0.97,[43] compared to the present UB-ST for skiers. Although the relative reliability observed for the ski-specific explosive upper-body strength and power measured during the UB-ST could be categorized as excellent (ICC >0.90), the absolute reliability measures (CV = 4.4–7.8%) might be too low to implement this test for elite skiers. One explanation for the high ICC in the present study is the heterogeneous performance level in the sample. As the ICC is a ratio of the between-participant variation to the within-participant variation, sample heterogeneity can inflate it.[32] Therefore, the high ICC for the UB-ST performance variables should be interpreted with caution.

4.2.3 | SkiErg peak power test

The present study demonstrated good-to-excellent reliability for the SkiErg P_{peak} (CV = 2.4%, ICC = 0.99). Similar reliability values have been found for peak power measurements on a Concept2 rowing ergometer using a seven-stroke maximum effort test (CV = 2.2%; ICC = 0.96) in national-level rowers[44] and a 12-stroke maximum effort test at different resistance settings (CV = 2.6%–6.5%; ICC = 0.87–0.98) in physically active individuals.[23] A known issue with the Concept2 SkiErg is the inability of the ergometer performance monitor to provide valid power output measures at the very high stroke rate seen in sprint skiing. In the present study, corrupt data were the case in 10 out of 108 trials, where two to seven strokes of the power recording during the trial were discarded. Nevertheless, the present results suggest that the commercially available Concept2 ski ergometer provides a reliable assessment of peak power with sufficient ski specificity. When using ski ergometers for testing purposes, practitioners should ensure the best possible transfer to skiing on snow by facilitating proper skiing technique (e.g., use of elastic bands around the hip to allow dynamic forward lean and limits on the pulling length) and being aware of the differences in the direction of the pull between the force vectors observed on the ergometer and the dynamic force vectors when pushing with ski poles.

4.3 | Limitations

The present study has several limitations. First, significant differences between trials were found in the GXT test performance, LT_1 BLa, LT_2 performance, LT_2 BLa, GXT HR_{peak}, 24-min DP performance, $Sprint_{1\text{-}min}$ HR_{peak}, $Sprint_{1\text{-}min}$ BLa, $Sprint_{DP2}$ V_{peak}, and UB-ST MIT P_{peak} (all $p < 0.05$), suggesting the presence of a systematic bias. This bias could stem from influencing factors such as learning effects (complex technical movements in skiing or the challenge of a pacing strategy during time trials), motivation, or fatigue.[12] The learning effect across skiers was surprising as all the participating skiers could be considered experienced, and all completed at least one previous familiarization trial for all test forms. Interestingly, significant performance increases between trials were observed in some of the most experienced skiers who had completed the particular test protocols >10 times in their athletic careers. This systematic bias from test to retest trial following a preceding familiarization trial was observed in similar studies for 2000-m time trial performance in well-trained rowers[41] and stroke rate in a seven-stroke test in national-level rowers.[44] At the same time, the improvements observed in some of the tests in the present study suggest that residual fatigue from the first trial was not an issue after the ≥48-h break between the test and retest trials. Further familiarization trials may be necessary for future investigations to minimize the bias, particularly in test protocols where adequate participant blinding is not possible, and test performance is significantly influenced by pacing. Caution is advised concerning the reliability of performance at LT_1 and LT_2, highlighting the difficulties of accurately separating the light- and medium-intensity domains using the described methodology of threshold determination. Another limitation concerns the sample of female and male junior and senior skiers with different testing experiences and performance levels. The heterogeneity of the study sample might have inflated the ICC. Testing elite athletes exclusively would likely result in lower ICC and higher CV values. In the context of performance changes in elite athletes, we consider CV the more relevant reliability estimate. Nevertheless, the sample of skiers for both measurement periods ($n = 27$), comprising men and women, was considerably larger

than the sample size used in the reliability studies with skiers mentioned previously (all $n < 12$).

5 | CONCLUSION

XC skiing–specific performance indicators, such as $\dot{V}O_{2max}$, performance at LT_2, and 24-min DP performance, can be assessed with high test–retest reliability. At the same time, the performance at LT_1, LT_1 BLa, and LT_2 BLa demonstrated only poor-to-good reliability. DP sprint performance, 1-min self-paced skating sprint performance, and upper-body peak power demonstrated high reliability in the sprint and neuromuscular domains. In contrast, one-repetition explosive upper-body strength and power demonstrated poor-to-moderate absolute reliability. Previous studies focused mainly on sprint test modalities[8,9] or only ergometer performance,[15–17] while this study provides test–retest reliability statistics and MDC estimates for comprehensive laboratory test procedures in XC skiing covering the aerobic, sprint, and neuromuscular performance domains. The findings suggest that most of the investigated ski-specific laboratory-derived sprint and endurance performance indicators are reliable and suitable for routine testing of XC skiers.

6 | PERSPECTIVE

This study provides CV and MDC measures across aerobic, sprint, and neuromuscular test performance and physiological characteristics in highly trained XC skiers. These reliability measures can help practitioners distinguish a genuine change in performance from measurement variability in within- and between-season measurements. As numerous performance protocols for XC skiing have been presented in the literature over the past 25 years, more studies investigating the reliability measures are needed to evaluate a test's quality and interpret laboratory test results from intervention studies. Considering the fact that (1) the smallest worthwhile enhancement in elite XC skiers is <1%,[11] and (2) laboratory testing is the currently available option to objectively assess XC ski-specific performance, testing facilities should establish the precision of measurement for each test protocol to increase the quality of the test interpretation. Future investigations should determine how the presented laboratory-based test protocols and corresponding performance indicators predict off- and on-snow sprint and distance XC skiing performance to improve the coaches' and athletes' ability to optimize the design and systematic application of different tests during the training process. Using the presented CV and MDC estimates, future studies may examine the long-term development of key performance indicators in XC skiers.

ACKNOWLEDGMENTS

The authors like to express their gratitude to all the athletes who participated in the study. In addition, the authors also like to thank Thomas Blokker, Stefan Ulrich, Carlotta Croci-Maspoli, and Bastien Krumm for their assistance with the data collection. Open access funding provided by Universite de Lausanne.

FUNDING INFORMATION

The Swiss Federal Institute of Sport, Magglingen, supported this study.

CONFLICT OF INTEREST STATEMENT

The authors do not have a conflict of interest.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study are accessible upon request from the corresponding author.

ORCID

Elias Bucher https://orcid.org/0000-0002-3689-2974
Grégoire P. Millet https://orcid.org/0000-0001-8081-4423
Jon P. Wehrlin https://orcid.org/0000-0002-8854-7937
Thomas Steiner https://orcid.org/0000-0002-3809-1230

REFERENCES

1. Svensson D, Sörlin S. The 'physiologization' of skiing: the lab as an obligatory passage point for elite athletes? *Sport Soc*. 2018;22(9):1574-1588. doi:10.1080/17430437.2018.1435031
2. Currell K, Jeukendrup AE. Validity, reliability and sensitivity of measures of sporting performance. *Sports Med*. 2008;38(4):297-316. doi:10.2165/00007256-200838040-00003
3. Carlsson M, Carlsson T, Hammarstrom D, Tiivel T, Malm C, Tonkonogi M. Validation of physiological tests in relation to competitive performances in elite male distance cross-country skiing. *J Strength Cond Res*. 2012;26(6):1496-1504. doi:10.1519/JSC.0b013e318231a799
4. Carlsson M, Carlsson T, Wedholm L, Nilsson M, Malm C, Tonkonogi M. Physiological demands of competitive Sprint and distance performance in elite female cross-country skiing. *J Strength Cond Res*. 2016;30(8):2138-2144. doi:10.1519/JSC.0000000000001327
5. Sandbakk O, Ettema G, Leirdal S, Jakobsen V, Holmberg HC. Analysis of a sprint ski race and associated laboratory determinants of world-class performance. *Eur J Appl Physiol*. 2011;111(6):947-957. doi:10.1007/s00421-010-1719-9
6. Losnegard T, Myklebust H, Hallen J. Anaerobic capacity as a determinant of performance in sprint skiing.

Med Sci Sports Exerc. 2012;44(4):673-681. doi:10.1249/MSS.0b013e3182388684
7. Sollie O, Losnegard T. Anthropometrical and physiological determinants of laboratory and on-snow performance in competitive adolescent cross-country skiers. *Front Physiol*. 2022;13:819979. doi:10.3389/fphys.2022.819979
8. Stoggl T, Lindinger S, Muller E. Reliability and validity of test concepts for the cross-country skiing sprint. *Med Sci Sports Exerc*. 2006;38(3):586-591. doi:10.1249/01.mss.0000190789.46685.22
9. Stoggl T, Lindinger S, Muller E. Evaluation of an upper-body strength test for the cross-country skiing sprint. *Med Sci Sports Exerc*. 2007;39(7):1160-1169. doi:10.1249/mss.0b013e3180537201
10. Losnegard T, Myklebust H, Spencer M, Hallen J. Seasonal variations in VO2max, O_2-cost, O_2-deficit, and performance in elite cross-country skiers. *J Strength Cond Res*. 2013;27(7):1780-1790. doi:10.1519/JSC.0b013e31827368f6
11. Spencer M, Losnegard T, Hallen J, Hopkins WG. Variability and predictability of performance times of elite cross-country skiers. *Int J Sports Physiol Perform*. 2014;9(1):5-11. doi:10.1123/ijspp.2012-0382
12. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med*. 1998;26(4):217-238. doi:10.2165/00007256-199826040-00002
13. Hopkins WG, Schabort EJ, Hawley JA. Reliability of power in physical performance tests. *Sports Med*. 2001;31(3):211-234. doi:10.2165/00007256-200131030-00005
14. Losnegard T, Andersen M, Spencer M, Hallen J. Effects of active versus passive recovery in sprint cross-country skiing. *Int J Sports Physiol Perform*. 2015;10(5):630-635. doi:10.1123/ijspp.2014-0218
15. Govus A, Marsland F, Martin D, Chapman D. Validity and reliability of an incremental double poling protocol in cross-country skiers. *J Hum Sport Exerc*. 2015;10(3):827-834. doi:10.14198/jhse.2015.103.08
16. Wisloff U, Helgerud J. Evaluation of a new upper body ergometer for cross-country skiers. *Med Sci Sports Exerc*. 1998;30(8):1314-1320. doi:10.1097/00005768-199808000-00021
17. Holmberg HC, Nilsson J. Reliability and validity of a new double poling ergometer for cross-country skiers. *J Sports Sci*. 2008;26(2):171-179. doi:10.1080/02640410701372685
18. Pellegrini B, Sandbakk Ø, Stöggl T, et al. Methodological guidelines designed to improve the quality of research on cross-country skiing. *J Sci Sport Exerc*. 2021;3(3):207-223. doi:10.1007/s42978-021-00112-6
19. McKay AKA, Stellingwerff T, Smith ES, et al. Defining training and performance caliber: a participant classification framework. *Int J Sports Physiol Perform*. 2022;17(2):317-331. doi:10.1123/ijspp.2021-0451
20. Howley ET, Bassett DR Jr, Welch HG. Criteria for maximal oxygen uptake: review and commentary. *Med Sci Sports Exerc*. 1995;27(9):1292-1301.
21. Borg G. Perceived exertion as an indicator of somatic stress. *Scand J Rehabil Med*. 1970;2(2):92-98.
22. Bishop D, Jenkins DG, Mackinnon LT. The relationship between plasma lactate parameters, Wpeak and 1-h cycling performance in women. *Med Sci Sports Exerc*. 1998;30(8):1270-1275. doi:10.1097/00005768-199808000-00014
23. Metikos B, Mikulic P, Sarabon N, Markovic G. Peak power output test on a rowing ergometer: a methodological study. *J Strength Cond Res*. 2015;29(10):2919-2925. doi:10.1519/JSC.0000000000000944
24. Hoffman MD, Clifford PS, Bota B, Mandli M, Jones GM. Influence of body mass on energy cost of roller skiing. *Int J Sport Biomech*. 1990;6(4):374-385. doi:10.1123/ijsb.6.4.374
25. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. L Erlbaum Associates; 1988:567.
26. Hopkins WG. Spreadsheets for analysis of validity and reliability. *Sportscience*. 2015;19:36-44.
27. Koo TK, Li MY. A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
28. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005;19(1):231-240. doi:10.1519/15184.1
29. Shechtman O. Chapter 4. The coefficient of variation as an index of measurement reliability. In: Doi SAR, Williams GM, eds. *Methods of Clinical Epidemiology*. Springer Series on Epidemiology and Public Health. Springer Berlin Heidelberg; 2013:39-49.
30. Turner A, Brazier J, Bishop C, Chavda S, Cree J, Read P. Data analysis for strength and conditioning coaches. *Strength Cond J*. 2015;37(1):76-83. doi:10.1519/ssc.0000000000000113
31. Ettema M, Brurok B, Baumgart JK. Test-retest reliability of physiological variables during submaximal seated upper-body poling in able-bodied participants. *Front Physiol*. 2021;12:749356. doi:10.3389/fphys.2021.749356
32. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med*. 2000;30(1):1-15. doi:10.2165/00007256-200030010-00001
33. R Core Team. *R: a language and environment for statistical computing*. R foundation for statistical computing. Version v 4.0.2. http://www.R-project.org/; 2021.
34. Cerezuela-Espejo V, Courel-Ibanez J, Moran-Navarro R, Martinez-Cava A, Pallares JG. The relationship between lactate and Ventilatory thresholds in runners: validity and reliability of exercise test performance parameters. *Front Physiol*. 2018;9:1320. doi:10.3389/fphys.2018.01320
35. Bjorklund G, Stoggl T, Holmberg HC. Biomechanically influenced differences in O2 extraction in diagonal skiing: arm versus leg. *Med Sci Sports Exerc*. 2010;42(10):1899-1908. doi:10.1249/MSS.0b013e3181da4339
36. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. *Med Sci Sports Exerc*. 1999;31(3):472-485. doi:10.1097/00005768-199903000-00018
37. Pallares JG, Moran-Navarro R, Ortega JF, Fernandez-Elias VE, Mora-Rodriguez R. Validity and reliability of Ventilatory and blood lactate thresholds in well-trained cyclists. *PloS One*. 2016;11(9):e0163389. doi:10.1371/journal.pone.0163389
38. Morton RH, Stannard SR, Kay B. Low reproducibility of many lactate markers during incremental cycle exercise. *Br J Sports Med*. 2012;46(1):64-69. doi:10.1136/bjsm.2010.076380
39. Palmer GS, Dennis SC, Noakes TD, Hawley JA. Assessment of the reproducibility of performance testing on an air-braked cycle ergometer. *Int J Sports Med*. 1996;17(4):293-298. doi:10.1055/s-2007-972849

40. Laursen PB, Francis GT, Abbiss CR, Newton MJ, Nosaka K. Reliability of time-to-exhaustion versus time-trial running tests in runners. *Med Sci Sports Exerc*. 2007;39(8):1374-1379. doi:10.1249/mss.0b013e31806010f5
41. Schabort EJ, Hawley JA, Hopkins WG, Blum H. High reliability of performance of well-trained rowers on a rowing ergometer. *J Sports Sci*. 1999;17(8):627-632. doi:10.1080/026404199365650
42. Stoggl T, Pellegrini B, Holmberg HC. Pacing and predictors of performance during cross-country skiing races: a systematic review. *J Sport Health Sci*. 2018;7(4):381-393. doi:10.1016/j.jshs.2018.09.005
43. Held S, Rappelt L, Donath L. Reliable peak power assessment during concentric and flexion-extension-cycle based rowing strokes using a non-modified rowing ergometer. *J Sports Sci Med*. 2022;21(1):131-136. doi:10.52082/jssm.2022.131
44. Nugent F, Comyns T, Chéilleachair N, Warrington G. Within-session and between-session reliability of the seven-stroke maximal effort test in National Level Senior Rowers. *J Australian Strength Cond*. 2019;27(4):22-28.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bucher E, Millet GP, Wehrlin JP, Steiner T. Test–retest reliability of ski-specific aerobic, sprint, and neuromuscular performance tests in highly trained cross-country skiers. *Scand J Med Sci Sports*. 2023;00:1-17. doi:10.1111/sms.14473