KAPITTEL 11

Prinsipper ved test–retest-reliabilitet

Test av protokoll for maksimal innsats på sykkelrulle med ett minutts varighet

Tommy Haugen[¥], Kjetil Marius Ulland Salvesen & Rune Høigaard

Universitetet i Agder, Fakultet for helse- og idrettsvitenskap

Sammendrag: Formålet med studien var todelt: 1) Gjennomgå sentrale prinsipper for test–retest-reliabilitet, og 2) reliabilitetsteste en protokoll for maksimal innsats på sykkelrulle med ett minutts varighet hos relativt godt trente unge voksne. Forsøkspersonene (N = 12) gjennomførte tre forsøk på sykkelrulle hvor gjennomsnittwatt ble registrert. Resultatene viste ingen signifikant forskjell mellom forsøkene, høy ICC (.996), lav CV og lav TE og SEM. Resultatene indikerte at sykkeltesten har evne til å produsere nøyaktige og stabile målinger. De indikerte også at det ikke ser ut til å forkomme større systematiske feilmålinger som påvirker resultatene i betydelig grad.

Nøkkelord: reliabilitetstest, sykkelergometer, innsats, prestasjon

Abstract: The study aim was twofold: 1) Review basic principles of test-retest reliability, and 2) investigate the reliability of a test procedure of a one-minute time trial on an ergometer bicycle in relatively well-trained young adults. The subjects (N = 12) completed three "all-out" intervals on the bicycle ergometers where average watt produced was recorded. The results of the experiments showed no significant mean differences across tests, high ICC (.996), low CV, TE, and SEM. The findings suggested that the ergometer test had the ability to produce accurate and stable measurements. It also indicated that measurement biases that affected the results did not seem to occur.

Keywords: reliability test, bicycle ergometer, effort, performance

[¥]Korresponderende forfatter: Tommy Haugen, Fakultet for helse- og idrettsvitenskap, Universitetet i Agder, Postboks 422, 4604 Kristiansand, tlf: +47 38 14 23 27, e-post: tommy.haugen@uia.no

Sitering av denne artikkelen: Haugen, T., Salvesen, K.M.U. & Høigaard, R. (2018). Prinsipper ved test–retest-reliabilitet: Test av protokoll for maksimal innsats på sykkelrulle med ett minutts varighet. I T. Haugen & R. Høigaard (red). *Trender i idrettspsykologisk forskning i Skandinavia* (Kap. 11, s. 225–238). Oslo: Cappelen Damm Akademisk. DOI: https://doi.org/10.23865/noasp.39.ch11

Introduksjon

Ifølge Murphy og Murphy (2012) legges det betydelig vekt på prestasjon og prestasjonsutvikling innenfor idrettspsykologisk forskning. Ønsket om å avdekke sosiale/psykologiske faktorer som kan påvirke idrettslige prestasjoner (kausalitet), får konsekvenser for valg av studiedesign. Dersom målet for forskningen er å etablere og demonstrere kausale effekter (årsak–virkning) mellom to eller flere variabler, er eksperimentelle studier å foretrekke (Gravetter & Forzano, 2016).

Idrettspsykologiske eksperimenter medfører behov for kontroll, og av den grunn foregår de ofte under strengt kontrollerte betingelser, som for eksempel i en laboratorium-kontekst. Haugen, Reinboth, Hetlelid, Peters & Høigaard (2016) og Nilsen, Haugen, Reinboth, Peters & Høigaard (2014) har for eksempel benyttet seg av repeterte målinger av prestasjon (innsats) på sykkelrulle i et fysiologisk testlaboratorium, hvor en rekke potensielt konfunderende variabler kan kontrolleres (f.eks. vær, temperatur og belastning). Likevel mangler de nevnte studiene en eksplisitt redegjørelse for reproduserbarheten for den benyttede protokollen. I et forskningsdesign som tar i bruk repeterte målinger, vil det alltid være en mulighet for at resultatene blir påvirket av tidligere forsøk (Hassmén & Hassmén, 2008). Grunner til dette kan være at gjennomføring av tester skaper en læringseffekt, eller at de bidrar til fysisk eller mental utmattelse. Det er derfor avgjørende at målemetodene har høy stabilitet, i tillegg til å kunne registrere naturlige variasjoner (sensitivitet) over tid (Portney & Watkins, 2000) eller endringer i betingelser. Det vil være avgjørende for studiens evne til å etablere kausale slutninger at eventuelle endringer i målte variabler kan relateres direkte til studiens manipulasjon. Innenfor metodisk litteratur kalles gjerne denne stabiliteten i målinger for reliabilitet. Et måleinstruments reliabilitet kan forklares som stabiliteten og konsistensen til målingene under like betingelser (Field, 2016).

Formålet med denne artikkelen er todelt: Å presentere en gjennomgang av sentrale prinsipper ved test–retest-reliabilitet, samt å reliabilitetsteste en protokoll for repetert maksimal innsats på sykkelrulle med ett minutts varighet, tidligere brukt i Haugen et al. (2016) og Nilsen et al. (2014).

Testing av reliabilitet

Reliabilitet refererer til nøyaktigheten og konsistensen til de dataene som samles inn ved en undersøkelse (Polit & Beck, 2010). Samme måling utført på samme person flere ganger og under samme betingelser skal i utgangspunktet gi samme verdi. Imidlertid er det uvanlig (om ikke umulig) å finne instrumenter som er feilfrie. Alle testmetoder er feilbarlige, og personer responderer med en eller annen grad av inkonsekvens (Field, 2016). Reliabilitet som begrep benyttes gjerne for å reflektere graden av feil (error) som enhver måling innehar (Streiner, Norman, & Cairney, 2014).

For å oppnå høy reliabilitet er en avhengig av at målingene er nøyaktige, og at målefeilene blir eliminert så langt det er mulig. Klassisk reliabilitetsteori tar utgangspunkt i at alle målinger, eller observert score, består av en sann score og en feilkomponent (error), der den sanne scorens verdi er uavhengig av målemetoden (Thomas, Nelson, & Silverman, 2005). Differansen mellom den sanne og den observerte scoren blir ofte betraktet som summen av tilfeldige og systematiske feilmålinger. Ved å identifisere graden av målefeil bedres forutsetningen for å forutse variasjon som forekommer ved målinger (Portney & Watkins, 2000). Tilfeldige feilmålinger (random error) påvirker resultatet av en test på en uforutsigbar måte. Disse målefeilene kan føre til tilfeldige resultater, og de kan påvirke ulikt fra måling til måling, og fra individ til individ (Weir, 2005). Årsaken til tilfeldige feilmålinger kan være biologiske og/eller mekaniske, eller de kan skyldes en ikke-standardisert testprotokoll. Ofte er komponentene av tilfeldige feil større enn komponentene av systematisk målefeil (Atkinson & Nevill, 1998).

Systematiske feil (bias/systematic error) har på sin side en tendens til å påvirke resultatet (kollektivt) i en bestemt retning. Det kan for eksempel forekomme ved tillæring av en bestemt oppgave som måles, ved utilstrekkelig restitusjon ved utmattende fysiske tester eller ved feil kalibrering av måleinstrument (Portney & Watkins, 2000). For å sikre at målingene ikke blir påvirket av tretthet, hukommelses- eller læringseffekt, er det avgjørende at tidsrommet mellom målingene blir tilpasset studien og metoden som benyttes (Robertson, Burnett, & Cochrane, 2014). Dersom intervallet mellom målingene blir for langt, kan det skje naturlige endringer i den målevariabelen man har til hensikt å påvirke (Hassmén & Hassmén, 2008; Robertson et al., 2014).

I denne sammenhengen er test–retest-reliabilitet av spesiell interesse (Rousson, Gasser, & Seifert, 2002). Dette kan gi et mål på instrumentets tidsmessige stabilitet (Hassmén & Hassmén, 2008). Ved en slik reliabilitetstesting blir ofte en gruppe individer utsatt for samme test/måling ved to eller flere anledninger, hvor en så sammenligner resultatene (Portney & Watkins, 2000). For å undersøke om det forekommer indikasjoner på systematiske feil (f.eks. læringseffekt, tretthet eller motivasjonssvikt) i en gruppe som måles ved flere anledninger, brukes gjerne paired sample t-test (to tester) eller repeated measures ANOVA (tre eller flere tester; Polit & Beck, 2010). Undersøkelse av eventuelle forskjeller mellom gjennomsnitt fra test til test gir svar på hvorvidt gruppens testresultater som helhet reduseres eller økes (Polit & Beck, 2010). En reliabel test uten betydelig systematisk feilmåling vil produsere lik sentraltendens for gruppen på tvers av tester.

Begrepet reliabilitet omfatter gjerne både relativ og absolutt reliabilitet (Baumgartner, 2003; Weir, 2005). Relativ reliabilitet er knyttet til graden av stabilitet i rank-posisjon i et utvalg, på tvers av gjentatte tester. Absolutt reliabilitet er på sin side knyttet til graden av individuell variasjon fra måling til måling (Domholdt, 2005).

Relativ reliabilitet kan kvantifiseres ved en korrelasjonskoeffisient (Atkinson & Nevill, 1998). Pearsons produkt-moment korrelasjonskoeffisient er en av de mest brukte estimeringene (ved to tester) av reproduserbarheten av målinger (Polit & Beck, 2010). En høy koeffisient viser at to ulike målinger gir relativt like resultater for de fleste individer. Dette vil indikere at testen gir stabile målinger over en viss tid (Polit & Beck, 2010). Ved en ustabil test, der målingene er tatt med kort tidsintervall (eller uten at noen forventede forandringer skal ha forekommet), vil korrelasjonskoeffisienten være lav. Korrelasjonskoeffisienten kan anses som et uttrykk for presisjonen av målingene, men kan påvirkes av homogenitet i utvalget (Hopkins, 2000). Ved tre eller flere gjentatte målinger på samme variabel benyttes gjerne en multivariat reliabilitetskoeffisient, for eksempel Intraclass correlation (ICC). Hva som kan anses som en akseptabel koeffisient, må vurderes på bakgrunn av hva som måles, måleinstrumentets nøyaktighet og hva resultatene skal benyttes til (Hassmén & Hassmén, 2008). Det vil si at hva som kan betraktes som tilfredsstillende

korrelasjon, vil variere. For tekniske måleinstrumenter forventes høyere koeffisient enn det som kreves ved måling av fenomener som i mindre grad lar seg påvirke av forsøkspersonene eller testleder (Hassmén & Hassmén, 2008). Ved kliniske forsøk er en ICC/*r*-verdi ~ .90 foreslått som en indikasjon på «høy» relativ reliabilitet (Hassmén & Hassmén, 2008). Gjennomføring av fysiske prestasjonstester på friske voksne i alderen 18 til 30 år hvor testen innehar veldefinerte prestasjonsscorer, tenderer til å gi reliabilitetsmålinger som ligger rundt øvre del av .70 til nedre del av .90 (Baumgartner, 2003).

Korrelasjon mellom to eller flere målinger sier noe om samvariasjonen mellom dem: Dersom den relative rangeringen i et utvalg er stabil, vil dette gi seg utslag i høy korrelasjon. Men høy korrelasjon kan ikke gi oss informasjon om systematiske feil, og det er anbefalt at relativ reliabilitet suppleres med mål på absolutt reliabilitet (Chinn, 1990; Domholdt, 2005). Lav grad av individuell variasjon vil tilsi høy grad av absolutt reliabilitet. For å måle den absolutte reliabiliteten kan eksempelvis Typical Error (TE; $SD_{diff}/\sqrt{2}$), Standard Error of Measurement (SEM; SDpooled*$\sqrt{(1-r)}$), variasjonskoeffisienten (CV; SD_{diff}/M_{pooled}*100) samt Bland-Altmans 95% limits of agreement (Bland & Altman, 1986) benyttes. TE og SEM kvantifiserer «typisk» målefeil i samme enhet som instrumentets måleenhet (Stratford & Goldsmith, 1997). Lavere verdier indikerer at målingsmetoden innehar høyere absolutt reliabilitet (Burton, Conway, & Holgate, 2000). Små individuelle endringer muliggjør oppdagelsen av små, men avgjørende endringer i målingsvariabelen (Batterham & George, 2003). CV oppgis i prosent og er en variasjonskoeffisient som er uavhengig av måleenhet (dimensjonsløs). Høyere reliabilitet vises ved lavere CV (Hopkins, 2000). Heterogenitet i utvalget har mindre påvirkning på TE, SEM og CV enn hva tilfellet er ved Pearsons korrelasjonskoeffisient (Burton et al., 2000).

Bland-Altmans 95% limits of agreement er et annet mål på absolutt reliabilitet, og innebærer som regel presentasjon av et såkalt Bland-Altman-plott, som visuelt illustrerer sammenhengen mellom to målinger. 95% limits of agreement indikerer et intervall for feilmålinger som lar seg evaluere med bakgrunn i praktisk relevans. Bland-Altman-plottet illustrerer gjennomsnittsverdien (x-akse) og differansen mellom to målinger (y-akse). Eventuelle uteliggere (outliers) og feil kan derfor observeres visuelt.

Reliabilitetstest av protokoll for repetert maksimal innsats på sykkelrulle med ett minutts varighet

Tidligere studier (Haugen et al., 2016; Nilsen et al., 2014) har i idrettspsykologiske eksperimenter som nevnt benyttet en protokoll for (repeterende) innsats/prestasjon på sykkelrulle under ulike sosiale/psykologiske betingelser, men protokollen er – så langt vi kjenner til – ikke tidligere reliabilitetstestet. Videre i denne studien ønsker vi av den grunn å reliabilitetsteste en protokoll for repetert maksimal innsats på sykkelrulle med ett minutts varighet.

Metode

Utvalg. Deltakerne i studien er et bekvemmelighetsutvalg bestående av bachelorstudenter i idrettsvitenskap ved Universitetet i Agder. Inklusjonskriteriene var 1) (selvvurdert) god fysisk form og 2) skade- og sykdomsfrihet. Åtte menn og fire kvinner med gjennomsnittlig alder 23.3 (*sd* = 1.4) år samtykket til deltakelse. Seks av deltakerne var aktive konkurranseidrettsutøvere (fotball eller håndball), de resterende var tidligere aktive idrettsutøvere. Gjennomsnittlig trente de 4.9 (*sd* = 0.40) økter per uke med en gjennomsnittlig varighet på 1.4 (*sd* = 0.5) timer per økt. Deltakerne ble i forkant av studien informert om studiens hensikt, risiko og dens frivillige karakter. Studien ble godkjent av Etisk komité ved Universitetet i Agder.

Prosedyrer

Forsøkspersonene gjennomgikk tre tester på sykkelrulle (ett minutt med maksimal innsats) under identiske betingelser: familiseringstest (Fam), test 1 (T1) og test 2 (T2). Testene ble gjennomført på samme dag med en total gjennomføringstid på omtrent fire timer. Deltakerne ble i forkant av sykkeltestene bedt om å fylle ut et spørreskjema slik at vi kunne innhente bakgrunnsinformasjon. De fikk en innføring i bruk av sykkelen og prosedyren som skulle gjennomføres. Syklene ble innstilt etter deltakernes kroppsbygning og preferanser, og de fikk ca. ti minutter til å bli fortrolige med syklene samt utføre finjustering av sete og gir. Forsøkspersonene måtte også angi det giret (utvekslingen) de ville benytte gjennom alle testene.

I forkant av testene gjennomgikk forsøkspersonene en selvstyrt oppvarming på ti minutter. Hver test hadde en varighet på ett (1) minutt. Forsøkspersonene ble informert om å holde en selvregulert maksimal innsats i hver test, og at gjennomsnittlig wattproduksjon i hver test ville bli registrert. Ved oppstart var pedalene i horisontal posisjon. Testleder startet testene med en nedtelling fra tre før startsignal ble gitt. Verbal tilbakemelding om gjenværende tid ble gitt hvert 15. sekund. Restitusjonstiden mellom hver test var 45 minutter, inkludert ny oppvarming.

Alle tester ble gjennomført på CompuTrainer Lab-ruller (RacerMate, 2016) med Nakamura 3.0 sykkelmodeller. Ergometerrullene er datastyrt gjennom programvaren RaceMate One, som automatisk regulerer motstanden ved hjelp av en elektronisk brems. Det er påvist at CompuTrainer Lab-rullene har en konstant belastning over et bredt belastningsspekter med en nøyaktighet på ± 2.5% (Racer Mate, Seattle, WA, USA). Før hver test ble det gjennomført en 15 minutters progressiv oppvarming (av utstyret; gjennomført av testledere) hvor belastningen gradvis ble økt fra 50 til 150 watt. Den progressive oppvarmingen klargjorde sykkelrullene gjennom å varme opp sykkeldekkene og ergometerrullen slik at en oppnådde den påkrevde driftstemperaturen for korrekte målinger. Deretter ble rullemotstanden kalibrert ved en roll-down resistance-prosedyre, hvor hjulmotstanden ble kalibrert til mellom 3.50 og 4.00 pund. Kalibreringsverdien ble lagret for den enkelte sykkel og benyttet av programvaren til RaceMate One for å beregne gjennomsnittswatt gjennom hver test.

Statistiske analyser

Statistiske analyser er gjennomført i SPSS, versjon 24.0, og GraphPad Prism, versjon 6.07. Preliminære analyser av normalfordeling ble gjennomført på alle testtidspunkter. D'Agostino & Pearsons omnibus normality test og visuell vurdering av Q-Q-plott, histogram og boxplott ble benyttet. For å undersøke systematiske målefeil mellom familiseringstest, test 1 og test 2 ble det gjennomført repeated measures ANOVA. Parvise test–retest-korrelasjoner av målingene ble undersøkt og vurdert ved bruk av Pearson produkt-moment korrelasjonskoeffisent (*r*), samt Intraclass

correlation (ICC) koeffisient (two-way mixed) for alle tre gjennomføringene. En p-verdi < .05 ble ansett som statistisk signifikant. Absolute agreement for målingene ble undersøkt ved bruk av variasjonskoeffisent (CV%), Standard Error of Measurement (SEM) og Typical Error (TE). Bland-Altman 95% limits of agreement og tilhørende plott ble gjennomført for å illustrere test–retest-stabilitet.

Resultater

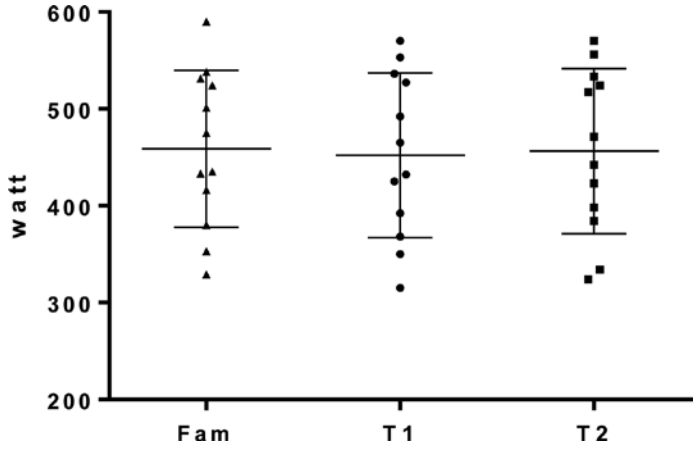
D'Agostino & Pearsons omnibus normality test viste tilfredsstillende normalfordeling på alle tre målepunkter (K2(p): Fam = 0.74(0.69); T1 = 1.74(0.42); T2 = 1.88(0.39)). Tabell 1 viser rådata for de tolv testpersonene for alle tre tester. Figur 1 presenter gjennomsnittwatt og standardavvik (*sd*) for de tre testene, samt individuelle mål som illustrasjon på spredning.

Tabell 1. Prestasjon (watt) på ett minutts maksimal innsats på sykkelrulle for hver testperson i alle testforsøk.

Person	Fam	Fam vs. T1	T1	T1 vs T2	T2	Fam vs. T2
1[a, d]	538		536		533	
2[a, c]	524		527		524	
3[b, c]	329		315		324	
4[b, d]	380		368		384	
5[b, d]	416		392		398	
6[b, c]	353		350		334	
7[a, d]	501		492		517	
8[a, d]	590		570		570	
9[a, c]	531		553		556	
10[a, d]	435		432		442	
11[a, c]	433		425		423	
12[a, c]	475		465		471	
Pearson's r		0.991**		0.992**		0.987**
TE (watt)		8.37		7.40		9.83
CV (%)		2.60		2.30		3.04
SEM (watt)		7.73		7.24		9.21

Note. Verdiene oppgitt som gjennomsnittswatt produsert. N = 12. [a]mann, [b]kvinne, [c]aktiv idrettsutøver, [d]ikke aktiv idrettsutøver. TE = Typical Error; CV = Coefficient of Variation; SEM = Standard Error of Measurement; ** = statistisk signifikant (p < .01).

Som vist i figur 1 var det ingen signifikant forskjell i wattproduksjon mellom de ulike testforsøkene ($F(df) = 2.17(2)$, $p = .16$). Konsistensen på målingene over tid var høy: Gjennomsnittlig ICC var .996 for de tre testene (Fam, T1 og T2), og Pearson-korrelasjon mellom testene var for alle parvise sammenhenger .99 ($p < .01$). I tillegg viste SEM, TE og CV til relativt lav grad av variasjon (se tabell 1).



Figur 1. Gjennomsnittswatt og standardavvik for familisering, test 1 og test 2.

Note. Fam = familiseringstest; T1 = test 1; T2 = test 2. Ingen statistisk signifikant forskjell mellom noen av testene (repeated measures ANOVA: (F (df) = 2.17 (2), p = .16).

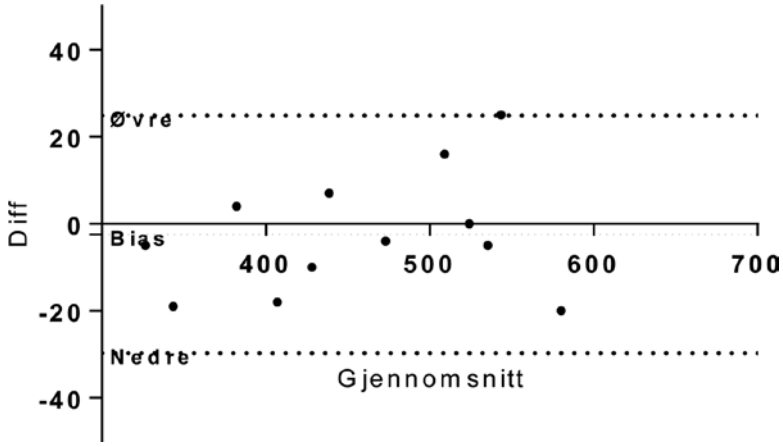
Figur 2–4 viser Bland-Altman-plott for målingene. Gjennomsnittlig bias for de tre parvise sammenhengene var -3.3.

Figur 2. Bland-Altman-plott for familisering vs. T1 test.

Note. Øvre = Upper 95% limit of agreement; Nedre = Lower 95% limit of agreement; Bias = Mean difference (6.67). Diff = differanse (T1-Fam).

Figur 3. Bland-Altman-plott for T1 vs. T2.

Note. Øvre = Upper 95% limit of agreement; Nedre = Lower 95% limit of agreement; Bias = Mean difference (–4.25). Diff = differanse (T2–T1).



Figur 4. Bland-Altman-plott for familisering vs. T2.

Note. Øvre = Upper 95% limit of agreement; Nedre = Lower 95% limit of agreement; Bias = Mean difference (–2.42). Diff = differanse (T2–Fam).

Diskusjon

Formålet med studien var å gjennomgå prinsipper for test–retest-reliabilitet, samt å reliabilitetsteste en protokoll for maksimal innsats på sykkelrulle med ett minutts varighet. Testprotokollen har blitt benyttet for å registrere (repeterende) innsats/prestasjon under ulike sosiale/

psykologiske betingelser (Haugen et al., 2016; Nilsen et al., 2014). Laboratorietester på innsats/prestasjon blir ofte brukt i idrettspsykologiske eksperimenter, hvor en empirisk forsøker å etterprøve kausale teoretiske mekanismer (se for eksempel Haugen et al., 2016; Høigaard, et al., 2006; Williams, et al., 1989). Det vil da være avgjørende at eventuell variasjon fra én test til en annen kan relateres til de manipulerte betingelsene, og ikke læringseffekt, tretthet eller tilfeldig målingsstøy. Reliabilitet er et komplekst begrep, og hvorvidt et måleinstrument er reliabelt eller ikke, er i stor grad åpent for fortolking.

Resultatene fra denne studien viser at den testede protokollen ser ut til å ha en tilfredsstillende evne til å produsere nøyaktige og stabile målinger. Det ser ikke ut til å forkomme systematiske feilmålinger som påvirker resultatene i betydelig grad. Funn viser med andre ord ingen tegn til markert læringseffekt (systematisk forbedret prestasjon fra én test til en annen) eller markert tretthet (systematisk reduksjon i prestasjon fra én test til en annen). Resultatene fra denne studien er i samsvar med lignende forsøk (for metastudie, se Hopkins, Schabert, & Hawley, 2001).

Imidlertid er det noen forhold rundt den aktuelle reliabilitetstestingen som kan være relevant å belyse. For det første var alle forsøkspersonene aktive utøvere eller tidligere aktive personer som anså seg selv å være i relativt god fysisk form. Hopkins et al. (2001) hevder at når en benytter godt trente deltakere, blir det gjerne høyere reliabilitet på powerbaserte tester enn når en benytter ikke-aktive personer. Trente individer har gjerne mer erfaring med høyintensivt arbeid gjennom trening og/eller konkurranse, og har dermed større evne til å opprettholde maksimal innsats. Innsatsperiodens lengde kan også påvirke individuell variasjon. Tester med en varighet på mindre enn 30 sekunder og på over 60 minutter er mindre stabile enn tester innenfor dette tidsspennet (Hopkins et al., 2001). Det forklares med at enkelte bevegelser blir mindre avgjørende for den totale prestasjonen dersom det er et høyt antall gjentakende bevegelser. Samtidig, med økende innsatsperiode vil intensitetsreguleringen og motivasjonen til å presse seg maksimalt kunne påvirkes og negativt påvirke reliabiliteten (Hopkins et al., 2001).

Restitusjonstiden mellom tester kan være avgjørende for resultatene. Ved et kort tidsrom mellom test og retest vil resultatene kunne variere på

grunn av tretthet, utmattelse eller utilstrekkelig næringsinntak (Hopkins et al., 2001). Derfor er det viktig å tilpasse restitusjonstiden mellom tester slik at forsøkspersonene er restituert og har mulighet til å prestere maksimalt ved repeterende forsøk. Rent fysiologisk kan en anta at 30 minutter eller mer er tilstrekkelig for fullstendig restitusjon etter 60 sekunders maksimal innsats (McArdle, Katch, & Katch, 2015). Resultatene fra reliabilitetsstudien indikerer relativt små forskjeller mellom testene, og 45 minutters restitusjon/pause mellom testene ser ut til å være tilfredsstillende for gjeldende prosedyre. En såpass lang restitusjonstid vil samtidig muliggjøre eksperimentell manipulasjon (i intervensjonsgruppe) i fremtidige studier, noe som kan være avgjørende for psykologiske eksperimenter (se for eksempel Nilsen et al., 2014).

Denne studien hadde som mål å undersøke relativ og absolutt reliabilitet i en protokoll med gjentatte prestasjonsmål i en kontrollert laboratoriekontekst. Hensikten var å etablere en trygghet om at fremtidige psykologiske eksperimenter kan gjennomføres, hvorpå eventuelle endringer kan tilskrives den aktuelle manipulasjonen. Imidlertid bør det legges til at studien ikke forholder seg til protokollens validitet. Validitet kan beskrives som en metodes evne til å måle det som det er til hensikt å måle (Polit & Beck, 2010; Portney & Watkins, 2000). Validiteten i en studie er blant annet avhengig av hvor relevante de empiriske dataene er for problemstillingen. På samme måte som reliabilitet er validitet avhengig av en kritisk evaluering av det metodiske grunnarbeidet (Polit & Beck, 2010). Høy grad av reliabilitet er en forutsetning for å oppnå høy validitet, men det er ikke en garanti for validiteten (Portney & Watkins, 2000).

Konklusjon

Resultatene i denne studien viser at en testprotokoll med selvstyrt maksimal innsats i ett minutt (på CompuTrainer Lab-ruller) produserer målinger med akseptabel nøyaktighet og stabilitet blant individer i relativt god fysisk form. Det kan derfor anses som en tilfredsstillende protokoll for bruk i eksperimentelle idrettspsykologiske studier som har til hensikt å manipulere på betingelser som kan påvirke individuell innsats. Imidlertid bør det presiseres at størrelsen på og homogeniteten (bachelorstudenter

i idrettsvitenskap) i utvalget medfører at resultatene ikke nødvendigvis kan generaliseres til en mer heterogen populasjon (utrente, eldre, barn), men først og fremst generaliseres til et lignende utvalg.

Referanser

Anshel, M.H. (1995). Examining social loafing among elite female rowers as a function of task duration and mood. *Journal of Sport Behavior, 18*, 39–50.

Atkinson, G. & Nevill, A.M. (1998). Statistical methods for assessing measurement error (reliability) in variable relevant to sports medicine. *Sports Medicine, 26,* 217–238.

Batterham, A.M. & George, K.P. (2003). Reliability in evidence-based clinical practice: A primer for allied health professionals. *Physical Therapy in Sport, 4,* 122–128.

Baumgartner, T.A. (2003). *Measurement for Evaluation in Physical Education and Exercise Science (7th ed).* Boston, MA: McGraw-Hill.

Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 8,* 307–310.

Burton, A., Conway, J.H., & Holgate, S.T. (2000) Reliability: What is it, and how is it measured? *Physiotherapy, 86,* 94–99.

Chinn, S. (1990). The assessment of methods of measurement. *Statistics in Medicine, 9*, 351–362.

Domholdt, E. (2005). *Rehabilitation research: principles and applications (3rd ed.).* St. Louis, MI: Elsevier Saunders.

Field, A. (2016). *An adventure in statistics: The reality enigma.* Thousand Oaks, CA: Sage Publications Inc.

Gravetter, F.J. & Forzano, L.-A.B. (2016). *Research Methods for the Behavioral Sciences (5th ed.)*. Stamford, CT: Cengage Learning.

Hassmén, N. & Hassmén, P. (2008). *Idrottsvetenskapliga forskningsmetoder.* Stockholm: SISU Idrottsböcker.

Haugen, T., Reinboth, M., Hetlelid, K.J., Peters, D.M., & Høigaard, R. (2016). Mental toughness moderates social loafing in cycle time-trial performance. *Research Quarterly for Exercise and Sport, 87*, 305–310.

Hopkins, W.G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine, 30*(1), 1–15.

Hopkins, W.G., Schabort, E.J., & Hawley, J.A. (2001). Reliability of power in physical performance tests. *Sports Medicine, 31*(3), 211–234.

Høigaard, R. (2010). *Social loafing in sport: From theory to practice*. Saarbrücken, Germany: VDM Verlag Dr. Müller Aktiengesellschaft & Co.

Høigaard, R., Tofteland, I., & Ommundsen, Y. (2006). The effect of team cohesion on social loafing in relay teams. *International Journal on Applied Sports Sciences, 18*, 59–73.

Karau, S.J. & Williams, K.D. (1993). Social loafing: A meta-analytic review and theoretical intergration. *Journal of Personality and Social Psychology*, 65, 681–706.

Latané, B. (1986). Responsibility and effort in organizations. I P. Goodman (red.), *Groups and organizations* (s. 277–303). San Francisco, CA: Jossey-Bass.

McArdle, W.D., Katch, F.I., & Katch, V.L. (2015). *Exercise Physiology: Nutrition, Energy, and Human Performance (8th ed.)*. Baltimore, MD: Wolters Kluwer, Lippincott Williams & Wilkins.

Polit, D.F. & Beck, C.T. (2010). *Essentials of Nursing Research: Appraising Evidence for Nursing Practice.* Philadelphia, PA: Wolters Kluwer, Lippincott Williams & Wilkins.

Portney, L.G. & Watkins, M.P. (2000). *Foundations of Clinical Research: Applications to Practice (2nd ed.).* New Jersey, NJ: Prentice Hall.

RacerMate. (2016). Compu Lab. Available from: http://www.racermateinc.com/computrainer/ hentet 7. mai 2016.

Robertson, S.J., Burnett, A.F., & Cochrane, J. (2014). Tests Examining Skill Outcomes in Sport: A Systematic Review of Measurement Properties and Feasibility. *Sports Medicine, 44*(4), 501–518.

Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Statistics in Medicine, 21*(22), 3431–3446.

Stratford, P.W. & Goldsmith, C.H. (1997). Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Physical Therapy, 77*(7), 745–750.

Streiner, D.L., Norman, G.R., & Cairney, J. (2014). *Health measurement scales: a practical guide to their development and use (5th ed)*. St.Ives, UK: Oxford University Press.

Swain, A. (1996). Social loafing and identifiability: The mediating role of achievement goal orientations. *Research Quarterly for Exercise and Sport*, *67*, 337–344.

Thomas, J.R., Nelson, J., & Silverman, S. (2005). *Research Methods in Physical Activity,(5th ed)*. Champaign, IL: Human Kinetics.

Weir, J.P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength & Conditioning Research, 19*(1), 231–240.

Williams, K.D., Nida, S.A., Baca, L.D., & Latané, B. (1989). Social loafing and swimming: Effects of identifiability on individual and relay performance of intercollegiate swimmers. *Basic and Applied Social Psychology, 10*, 73–81.