Artikel online veröffentlicht: 03.02.2021

Eine Alternative zur klassischen Testtheorie?

Eine exemplarische Anwendung der Generalisierbarkeitstheorie auf der Basis von Sekundärdaten

An Alternative to Classic Test Theory?

An Exemplary Application of Generalizability Theory-based on Secondary Data

Autoren

Iris Sterkele[1], Pierrette Baschung Pfister[1, 2], Ruud Knols[1, 2, 3], Eling D de Bruin[3, 4, 5]

Institute

1 Physiotherapie Ergotherapie USZ, Universitätsspital Zürich, Zürich, Schweiz
2 Forschungszentrum Physiotherapie Ergotherapie USZ, Universitätsspital Zürich, Zürich, Schweiz
3 Departement Gesundheitswissenschaften und Technologie, Institut für Bewegungswissenschaften und Sport, ETH Zürich, Zürich, Schweiz
4 Division of Physiotherapy, Department of Neurobiology, Care Science and Society, Karolinska Institute, Stockholm, Sweden
5 Research Line Functioning and Rehabilitation, Department of Epidemiology, School CAPHRI Care and Public Health Research Institute, Maastrich University, The Netherland

Schlüsselwörter

Generalisierbarkeitstheorie, Reliabilität, Generalisierbarkeitsstudie, Physiotherapie

Key words

generalizability theory, reliability, generalizability study, physiotherapy

eingereicht 17.06.2020
akzeptiert 03.09.2020
online publiziert 03.02.2021

Bibliografie

physioscience 2021; 17: 25–33
DOI 10.1055/a-1201-6872
ISSN 1860-3092
© 2021. Thieme. All rights reserved.
Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Korrespondenzadresse

Iris Sterkele
Gloriastrasse 25, 8091 Zürich, Schweiz
Iris.Sterkele@usz.ch

Zusätzliches Material finden Sie unter https://doi.org/10.1055/a-1201-6872

ZUSAMMENFASSUNG

Hintergrund Messungen bilden die Grundlage des wissenschaftsbasierten therapeutischen Vorgehens. Messfehler sind jedoch ein verbreitetes Problem. Um Fehlerquellen zu ermitteln, sind erweiterte statistische Modelle notwendig. Im Gegensatz zur klassischen Testtheorie ermöglicht die Generalisierbarkeitstheorie, mehrere Fehlerquellen gleichzeitig zu untersuchen.

Ziel Vergleich der klassischen Testtheorie und der Generalisierbarkeitstheorie anhand eines physiotherapeutischen Beispiels zur Reliabilitätsermittlung und deren Ergebnisse.

Methode Die Messeigenschaft „Reliabilität“ wurde anhand von Daten einer vorangegangenen Reliabilitätsstudie, in der die klassische Testtheorie zum Einsatz kam, untersucht. Dazu wurde exemplarisch eine Generalisierbarkeitsstudie durchgeführt.

Ergebnisse Die Variabilität ging hauptsächlich von den Probanden und der Probanden-Untersucher-Interaktion aus. Unabhängig von fixen oder zufälligen Facetten waren die Generalisierbarkeitskoeffizienten für alle Bedingungen (overall, inter-rater, intra-Messzeitpunkte) exzellent.

Schlussfolgerung Die Ergebnisse deuten darauf hin, dass die Generalisierbarkeitstheorie gegenüber der klassischen Testtheorie Vorteile hat. Diese ermöglicht es, einen wirkungsvollen und effizienten Einsatz von Messinstrumenten im klinischen Alltag zu ermitteln.

ABSTRACT

Background Applying repeated measurements, errors can arise from multiple sources, which affects the outcome. In order to determine these sources of error, an expanded statistical model like the generalizability theory framework is required. In contrast to the classic test theory, it enables the simultaneous investigation of several sources of error.

Objective To demonstrate an exemplary application of generalizability theory in a physiotherapy setting based on an isometric muscle strength test as well as the comparison of two different approaches, the classical test theory and the generalizability theory.

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

Method Examinizing the clinimetric property „Reliability" based on data of a previous reliability study in which classic test theory was used. For this purpose, an exemplary generalizability study was conducted.

Results Variances resulted mainly from subjects and interactions between subjects and raters. Generalizability coefficients were excellent for all conditions (overall, inter-rater, inter-time point), regardless of the type of facets (fix or random facets).

Conclusion The results of the exemplary generalizability study shows that generalizability theory has advantages over classic test theory. Generalizability theory offers an approach to determine and implement effective and efficient use of measurement protocols for physiotherapy practice.

Einleitung

Die Wirksamkeit therapeutischer Maßnahmen systematisch zu messen, ist ein wichtiger Teil der evidenzbasierten Praxis. Messungen und Assessments ermöglichen es, die Patientensituation ganzheitlich zu erfassen, einen Behandlungsplan zu erstellen und den Rehabilitationsverlauf zu überprüfen. Darüber hinaus dienen sie der Messbarkeit des Therapieerfolges und der Qualitätskontrolle. Messungen bilden auch die Grundlage des wissenschaftsbasierten therapeutischen Vorgehens. Behandlungseffekte systematisch zu erfassen, ermöglichen es, Guidelines zu erstellen, nicht ausreichend effektive Behandlungsmethoden zu eliminieren und die Kostenübernahme gegenüber den Kostenträgern wissenschaftlich zu begründen [1]. COSMIN (**CO**nsensus-basierte **S**tandards von gesundheitsbezogenen **M**ess**IN**strumenten) ist eine internationale, multidisziplinäre Forschungsgruppe, die consensus-basierte Standards zur Selektion gesundheitsspezifischer Messinstrumente entwickelt und evaluiert. Diese Gruppe definiert „Reliabilität" als das Ausmaß, in dem ein Messinstrument frei von Messfehlern ist [2]. Bei wiederholten Messungen gelangen die messenden Personen unter gleichen Bedingungen (Test-Retest) zum selben Ergebnis. Das Autorenteam unterscheidet zwei Fälle: (1) Der gleiche Untersucher wiederholt die Messung (Intraraterreliabilität) oder (2) ein anderer Untersucher wiederholt die Messung (Interraterreliabilität) [2]. Laut COSMIN-Checkliste umfasst die Domäne „Reliabilität" drei Messeigenschaften: „Reliabilität", „Messfehler" und „interne Konsistenz". Reliabilität ist die Fähigkeit eines Messinstruments, zwischen Probanden zu differenzieren. Messfehler ist ein Mass für die Präzision wiederholter Messungen durch einen oder mehrere Untersucher. Je kleiner der Messfehler, desto präziser das Messverfahren. Interne Konsistenz beschreibt das Ausmaß der Wechselbeziehung zwischen einzelnen Items. Sie ist ausschließlich für Fragebögen bedeutsam.

Um die Messeigenschaft Reliabilität von wiederholten Messungen zu erheben, wird in der Physiotherapie oft ein klassisches testtheoretisches Framework verwendet [2]. Bei der klassischen Testtheorie (KTT) setzt sich der beobachtete Wert (X) aus einem wahren Wert (η) und einem globalen Messfehler (E) zusammen. Die Messeigenschaft kann wie in Gleichung (1) dargestellt werden [2]:

$$X_h = \eta + E_h \tag{1}$$

Dabei steht h für wiederholte Messungen, welche entweder durch verschiedene Untersucher, bei verschiedenen Messgelegenheiten unter verschiedenen Umständen oder mit verschiedenen Messinstrumenten durchgeführt werden [2]. Der wahre Wert (η) wird als Durchschnittswert über unendlich viele Messungen gesehen. Dieser wird in Gleichung (2) dargestellt. Dabei ist (μ_x) der globale Durchschnitt über alle Beobachtungen (x) und (P_i) die personenspezifische Abweichung (i steht für den i-ten wahren Wert der i-ten Person aus einer Stichprobe mit Anzahl n Probanden).

$$\eta_i = \mu_x + P_i \qquad i = 1, 2, 3 \ldots n \tag{2}$$

Bei der KTT wird lediglich eine einzelne globale Fehlerquelle berücksichtigt. Es wird davon ausgegangen, dass der Messfehler und der wahre Wert unabhängig voneinander sind. Die KTT definiert die Reliabilität (R) als Verhältnis der Varianz des wahren Wertes (σ^2_η) durch die Varianz des beobachteten Wertes (σ^2_x) (Gleichung (3)) [2].

$$R = \frac{\text{Varianz des wahren Wertes } (\sigma^2_\eta)}{\text{Varianz des beobachtenden Wertes } (\sigma^2_X)} \tag{3}$$

Wie in Gleichung (1) beschrieben, setzt sich der beobachtete Wert (X) aus dem wahren Wert (η) und dem Messfehler (E) zusammen. Deshalb kann die Varianz des beobachteten Wertes (σ^2_x) mit der Summe aus der Varianz des wahren Wertes (σ^2_η) und der Fehlervarianz (σ^2_E) ersetzt werden. Daraus entsteht Gleichung (4) [2]:

$$R = \frac{\text{Varianz des wahren Wertes } (\sigma^2_\eta)}{\text{Varianz des wahren Wertes } (\sigma^2_X) + \text{Fehlervarianz } (\sigma^2_E)} \tag{4}$$

In der KTT erhält die Identifikation multipler Fehlerquellen keine Beachtung. Reliabilität kommt in Form des Intraklassenkorrelationskoeffizienten (ICC) zum Ausdruck [3, 4] und ist nur begrenzt aussagekräftig [5].

In der Physiotherapie erfolgen Messungen jedoch nicht unter vollständig kontrollierbaren Bedingungen. Aus diesem Grund können bei wiederholten Messungen Messfehler durch eine theoretisch unendlich große Vielzahl von Ursachen entstehen. Zu diesen Ursachen zählen beispielsweise die messende Person, das Messinstrument, das Messprotokoll, die Anzahl gewählter Messwiederholungen, die Bedingungen einer Messung oder die Probanden.

Um multiplen Messfehlerquellen in der klinischen Praxis gerecht zu werden, ist ein erweitertes statistisches Modell gefragt. Hier setzt die Generalisierbarkeitstheorie (G-Theorie) an [6, 7].

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

Bis heute wird die G-Theorie in der Physiotherapie relativ wenig angewendet. Eine Literatursuche in der Datenbank PubMed im März 2020 zur G-Theorie in der Physiotherapie ergab knapp 40 Treffer, eine Suche zur Reliabilität in der Physiotherapie hingegen über 6000 Treffer. Dabei wurden folgende Suchstrategien verwendet: Suche zur G-Studie: (Generalizability Theory OR G-Theory OR G-Study) AND (Physical therapy OR Physiotherapy); Suche zur Reliabilität: (Reliability OR intrarater Reliability OR interrater Reliability) AND (Physical therapy OR Physiotherapy).

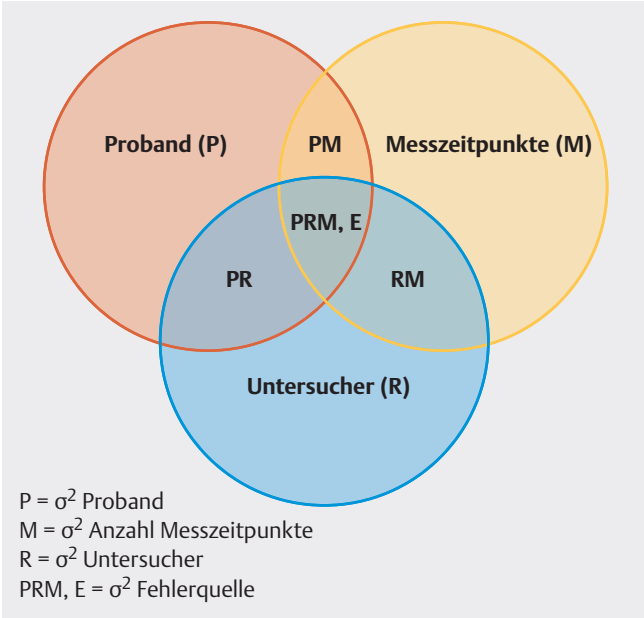
Das Ziel dieser Arbeit ist, eine Einführung in die Grundelemente der G-Theorie zu vermitteln. Anhand eines physiotherapeutischen Beispiels zur Reliabilitätsermittlung werden die KTT und die G-Theorie hinsichtlich der Ergebnisse miteinander verglichen. Als Beispiel dient eine typische Untersuchungsanlage. Die Forschungsfrage lautet: Unterscheiden sich der Intraklassenkorrelationskoeffizient (ICC) und die Generalisierbarkeitskoeffizienten (G-Koeffizienten) voneinander?

Theoretische Grundlagen der G-Theorie

Die G-Theorie stellt eine Erweiterung der KTT dar, indem sie den globalen Messfehler in einzelne Bestandteile zerlegt. Damit kann der Einfluss von mehreren Fehlerquellen und deren Interaktionen für ein beliebiges Messverfahren untersucht werden und es können verlässlichere Informationen über die Messzuverlässigkeit in der klinischen Praxis erzielt werden [2, 5, 8]. Die G-Theorie unterscheidet zwei Studientypen: G-Studie (Generalisierbarkeitsstudie) und D-Studie (Entscheidungsstudie) [9, 10]. In einer G-Studie werden für alle mit der Messung verbundenen Fehlerquellen bzw. Facetten und für deren Interaktionen der Beitrag an der Varianz untersucht und quantifiziert. Diese Varianzquellen werden verwendet, um zu bestimmen, welche Facetten oder welche Interaktionen am meisten zum Messfehler beitragen und folglich in klinischen Settings kontrolliert werden sollten. Mittels einer D-Studie wird untersucht, welche Messprotokolle in der klinischen Praxis optimal sind, um zuverlässige Daten zu erhalten [5].

In einem ersten Schritt wird das sogenannte Universum von zulässigen Beobachtungen (universe of admissible observations) definiert. In der G-Theorie wird unter dem Begriff Universum nicht der gesamte Kosmos verstanden, sondern die Bedingungen und Bestandteile (Facetten), welche das zu untersuchende Messverfahren beinhaltet [8]. Diese werden aufgrund theoretischer Vorüberlegungen vom Forscher definiert. Somit werden alle diejenigen Facetten eruiert, welche als Fehlerquelle im definierten Universum in Frage kommen. Sie gelten nur für das vom Forscher festgelegte Szenario [8].

Weiter muss das Messprozedere, das sogenannte Design, bestimmt werden. Dabei werden zwei Designs unterschieden, ein vollständig gekreuztes (crossed) oder ein verschachteltes (nested) Design. Bei einem vollständig gekreuzten Design werden alle denkbaren Kombinationen aus den verschiedenen Facetten berücksichtigt. Beispielsweise werden alle Probanden von allen Untersuchern zu jedem Messzeitpunkt gemessen. Bei einem verschachtelten Design findet eine Zuteilung statt. Die Probanden werden beispielsweise Untersucher A oder Untersucher B zugewiesen, sodass nicht jeder Untersucher alle Probanden misst [5]. Wird beispielsweise die Muskelkraft einer Stichprobe zu verschiedenen Zeitpunkten von unterschiedlichen Untersuchern gemessen, wobei jeder Untersucher jeweils alle Probanden zu jedem Messzeitpunkt misst, liegt ein vollständig gekreuztes zweifaktorielles Design vor (▶ **Abb. 1**).



▶ **Abb. 1** Venn-Diagramm eines zweifaktoriellen Designs (P × M × R × PRM, E). (Quelle: I. Sterkele, graf. Umsetzung Thieme Group).

Die Facetten bei diesem Messprozedere sind: Messzeitpunkt, Proband und Untersucher. Jeder beobachtete Wert (X_{ijk}) einer einzelnen Kraftmessung eines Probanden, erhoben durch einen einzelnen Untersucher, kann dann mit folgendem Modell (5) dargestellt werden:

$$X_{ijk} = \mu + P_i + R_j + M_k + PM_{ik} + PR_{ij} + RM_{jk} + PRM_{ijk} \quad (5)$$

$$i = 1,2,3 \ldots I; \quad j = 1,2,3 \ldots J; \quad k = 1,2,3 \ldots K$$

Dabei wird jede Kraftmessung durch 3 Facetten, Messzeitpunkt (M_k), Tagesform des Probanden (P_i) und des Untersuchers (R_j) und deren Interaktionen (PM_{ik}, PR_{ij}, RM_{jk}) zueinander sowie eines Residualfehlers (PRM_{ijk}), auch (E_{ijk}) genannt, beeinflusst (▶ **Tab. 1**).

In der G-Theorie werden die Facetten des definierten Universums von zulässigen Beobachtungen unterschiedlich bezeichnet. Die Facette, die sich auf das Messobjekt bezieht, heißt Facette der Differenzierung (D-Facette, facet of differentiation) [5, 8]. Häufig handelt es sich dabei um Patienten. Facetten können jedoch, abhängig vom Messkontext, auch Situationen, Dinge oder Szenarien sein. In jeder Studie existiert jeweils nur eine D-Facette [5, 8]. Die Facetten, die Fehlerquellen darstellen, werden Facetten der Generalisierung (G-Facette, facet of generalisation) genannt. Je nach Forschungsfrage können eine oder mehrere G-Facetten definiert werden [5, 8]. Ist die G-Facette beispielsweise die Facette „Messzeitpunkt“ (M), wird untersucht, inwieweit bei wiederholten Messungen von einem Messzeitpunkt auf einen anderen generali-

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

▶ **Tab. 1** Begriffserklärungen zu Modell (5).

X_{ijk}	**Beobachtungswert des i-ten Probanden, j-ten Untersuchers, k-ten Messzeitpunktes**
μ	globaler Durchschnitt aller Facetten, Interaktionen und des Residualfehlers
Facetten	
P_i	Effekt des i-ten Probanden
R_j	Effekt des j-ten Untersuchers
M_k	Effekt des k-ten Messzeitpunktes
Interaktionen	
PM_{ik}	Interaktionseffekt zwischen i-ten Probanden und k-ten Messzeitpunktes
PR_{ij}	Interaktionseffekt zwischen i-ten Probanden und j-ten Untersuchers
RM_{jk}	Interaktionseffekt zwischen j-ten Untersuchers und k-ten Messzeitpunktes
Residualfehler	
PRM_{ijk}	Residualfehler des i-ten Probanden, j-ten Untersuchers, k-ten Messzeitpunktes

siert werden kann. Wenn für 2 Facetten generalisiert wird, beispielsweise „Untersucher“ (R) und „Messzeitpunkt“ (M), wird untersucht, inwieweit die Messungen Messzeitpunkt und Untersucher unabhängig sind. Weiter wird unterschieden, ob eine Facette „fix“ (fix facets) ist, also ob sie im gesamten Messverfahren konstant bleibt und somit nicht fehlerbehaftet ist, oder ob sie „zufällig“ ist. Zufällige Facetten (random facets) sind diejenigen Facetten, die zum relevanten Fehler beitragen [5]. Im nächsten Schritt werden aus allen Facetten und deren Interaktionen des Modells Varianzkomponenten gebildet. Daraus wird anschließend der sogenannte Generalisierbarkeitskoeffizient (G-Koeffizient oder G) berechnet.

Jeder G-Koeffizient ist eine Erweiterung des klassischen Reliabilitätskoeffizienten. Analog zur Gleichung (4) stellt der G-Koeffizient ein Verhältnis aus der Varianz des universalen Wertes und der Summe der Varianz des universalen Wertes und der Fehlervarianz dar (Gleichung (6)) [8].

$$\text{G-Koeffizient} = \frac{\text{Varianz des universalen Wertes}}{\text{Varianz des universalen Wertes + Fehlervarianz}} \tag{6}$$

Der universale Wert (τ) ist der Erwartungswert (Durchschnitt) für einen Probanden über alle Facetten der Generalisation. Er entspricht dem „wahren“ Wert (η) der KTT. Im Gegensatz zur KTT besteht jedoch nicht ein einziger wahrer Wert, sondern abhängig davon, auf welche Facette oder Facetten des Modells generalisiert wird, verschiedene „wahre“ Werte [5, 8].

Je nach Messszenario, also abhängig davon, ob fixe oder zufällige Facetten vorhanden sind, wird der G-Koeffizient mit einer relativen Fehlervarianz (σ^2_δ) oder absoluten Fehlervarianz (σ^2_Δ) berechnet [5, 8]. Anhand von zwei unterschiedlichen Messszenarien

▶ **Tab. 2** Begriffserklärungen zu Gleichung (7).

Interaktion	
σ^2_{PR}	Varianz der Interaktion zwischen Probanden und Untersucher
Residualfehler	
σ^2_{PMR}	Varianz des Residualfehlers

wird mit dem vorgestellten Beispiel der Kraftmessung nachfolgend erläutert, bei welchem Messszenario die relative respektive absolute Fehlervarianz verwendet wird. In beiden Messszenarien ist die Facette, auf welche generalisiert wird, die Facette „Untersucher“ (R). In einem ersten Szenario werden die Messungen jeweils von denselben Untersuchern durchgeführt. Die Facette „Untersucher“ ist somit eine „fixe“ Facette. Da fixe Facetten nicht zum relevanten Fehler beitragen, wird davon ausgegangen, dass die Messung in diesem Fall nicht für systematische Verzerrungen durch die Untersucher korrigiert werden muss [5]. In diesem Messszenario besteht die Fehlervarianz lediglich aus der Summe der Varianz der Interaktionen zwischen Untersucher und Proband und einem Residualfehler (Gleichung 7). Diese wird relative Fehlervarianz genannt. Ein G-Koeffizient, welcher mit der relativen Fehlervarianz, also fixen Facetten, berechnet wird, wird mit dem Symbol $E\rho^2$ beschriftet (Gleichung 7.1) [8] (▶ **Tab. 2**).

$$\sigma^2_\delta = \sigma^2_{PR} + \sigma^2_{PMR} \tag{7}$$

$$\text{G-Koeffizient: } E\rho^2 = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\delta} \tag{7.1}$$

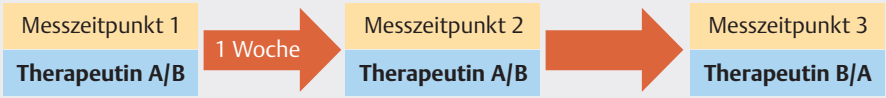
Im zweiten Szenario sind die Untersucher aus einer Zufallsstichprobe aller verfügbaren Untersucher zufällig gewählt. Die Facette „Untersucher“ ist damit eine „zufällige“ Facette. Zufällige Facetten sind fehlerbehaftet und tragen zum relevanten Fehler bei. Deshalb wird das Ausmaß der Verzerrung, das durch die Untersucher entsteht, in der Fehlervarianz berücksichtigt. Analog zum G-Koeffizient $E\rho^2$ wird der G-Koeffizient mit der absoluten Fehlervarianz σ^2_Δ berechnet und mit Φ angegeben. Dieser G-Koeffizient wird auch Zuverlässigkeitskoeffizient oder Index der Zuverlässigkeit genannt (index of dependability) (Gleichung 8) [8]).

$$\text{G-Koeffizient: } \Phi = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\Delta} \tag{8}$$

Die absolute Fehlervarianz besteht in diesem Beispiel aus der Summe aller Varianzkomponenten und Interaktionen mit Untersucher (R) und einem Residualfehler (Gleichung 8.1) (▶ **Tab. 3**).

$$\sigma^2_\Delta = \sigma^2_R + \sigma^2_{PR} + \sigma^2_{MR} + \sigma^2_{PMR} \tag{8.1}$$

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.



▶ **Abb. 2** Studienprozedere. (Quelle: I. Sterkele, graf. Umsetzung Thieme Group).

▶ **Tab. 3** Begriffserklärungen zu Gleichung (8.1).

Facette	
σ^2_R	Varianz des Untersuchers
Interaktionen	
σ^2_{MR}	Varianz der Interaktion zwischen Messzeitpunkt und Untersucher
σ^2_{PR}	Varianz der Interaktion zwischen Probanden und Untersucher
Residualfehler	
σ^2_{PMR}	Varianz des Residualfehlers

Für die Interpretation der G-Koeffizienten wird analog zur Interpretation der Reliabilitäts-Koeffizienten ein Kontinuum von 0–1 verwendet. Werte von < 0,4 bedeuten schlechte, 0,4–0,75 faire bis gute und > 0,75 exzellente Reliabilität [11].

Methode

Probanden und Prozedere

Für die vorliegende Studie wurden Daten von 50 Personen mit inflammatorischer Myopathie (IM) verwendet. Die Daten wurden im Rahmen einer vorangegangenen Reliabilitätsstudie erhoben, wobei die KTT für die Analyse eingesetzt wurde [12]. Die Messungen wurden von 2 erfahrenen Physiotherapeutinnen durchgeführt. Die isometrische Maximalkraft der Schulterabduktoren wurde mit dem MicroFET2, einem handgehaltenen Dynamometer (Hoggan Health Industries Inc. West Draper: Messeinheit: Newton (N)), zu 3 verschiedenen Messzeitpunkten insgesamt 3 × gemessen. Die erste Messung diente als Testdurchlauf. Aus der zweiten und dritten Messung wurde die durchschnittliche isometrische Maximalkraft für jeden Probanden berechnet. 21 Probanden wurden von Therapeutin A und 29 Probanden von Therapeutin B gemessen. Nach einer Woche testeten Therapeutinnen A und B dieselben Probanden erneut. Zum dritten Messzeitpunkt wechselten die Therapeutinnen (▶ **Abb. 2**). Eine ausführliche Beschreibung der Probanden, der Rekrutierung und des Studienprozederes erfolgte in [12].

Studiendesign

Aus dem Messkontext ergaben sich theoretisch die Facetten Proband (P_i), Messzeitpunkt (M_k), Rater (R_j) und die Interaktionen (PM_{ik}), (RM_{jk}), (PR_{ij}) sowie der Residualfehler (PMR_{ijk}). Idealerweise hätte jede Therapeutin alle Probanden zu 3 Messzeitpunkten

▶ **Tab. 4** Begriffserklärung zu Modell (9).

X_{ijk}	**Beobachtungswert des i-ten Probanden, j-ten Untersuchers, k-ten Messzeitpunktes**
μ	globaler Durchschnitt aller Facetten, Interaktionen und des Residualfehlers
Facetten	
P_i	Effekt des i-ten Probanden
R_j	Effekt des j-ten Untersuchers
M_k	Effekt des k-ten Messzeitpunktes
Interaktionen	
PR_{ij}	Interaktionseffekt zwischen i-ten Probanden und j-ten Untersuchers
RM_{jk}	Interaktionseffekt zwischen j-ten Untersuchers und k-ten Messzeitpunktes
Residualfehler	
PRM_{ijk}	Residualfehler

getestet – im Sinne eines vollständig gekreuzten zweifaktoriellen Designs (crossed two-factor design). Das konkrete Studienprozedere sah jedoch vor, dass zum Messzeitpunkt 1 eine zufällige Zuteilung der Probanden zu Therapeutin A oder Therapeutin B erfolgte. Dadurch wurden die Probanden zum Messzeitpunkt 1 lediglich von einer Untersucherin gemessen, wodurch die Interaktion PM_{ik} mit jeweils nur einem Messwert gemessen wurde. Das entspricht einem nicht vollständig gekreuzten Design mit zwei Facetten (Messzeitpunkt und Untersucher), weshalb nur die Haupteffekte (Effekt der Facetten) und 2 der 3 Interaktionseffekte geschätzt wurden. In diesem Modell (9) besteht der beobachtete Wert (X_{ijk}) aus dem globalen Durchschnitt (μ) aller abschätzbarer Facetten und Interaktionen sowie aus einem Residualfehler (▶ **Tab. 4**).

$$X_{ijk} = \mu + P_i + R_j + M_k + PR_{ij} + RM_{jk} + PRM_{ijk} \quad (9)$$
$$i = 1,2,3 \ldots 50; \quad j = 1,2; \quad k = 1,2,3$$

Im beschriebenen Modell waren die Probanden (P) die D-Facette. Exemplarisch wurden verschiedene G-Facetten definiert:

- (1) generalisieren für Untersucher (R) und Messzeitpunkt (M) G-overall
- (2) generalisieren für Untersucher (R) G-inter-rater
- (3) generalisieren für Messzeitpunkt (M) G-inter-Messzeitpunkt

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

▶ **Tab. 5** Varianzkomponenten aller Haupteffekte, Interaktionseffekte und des Residuals.

Varianzkomponenten der Haupteffekte	Beschreibung
σ^2_P	Varianzkomponente des Probanden
σ^2_R	Varianzkomponente des Untersuchers
σ^2_M	Varianzkomponente des Messzeitpunktes
Varianzkomponenten der Interaktionseffekte	
σ^2_{PR}	Varianzkomponente der Interaktion zwischen Probanden und Untersucher
σ^2_{RM}	Varianzkomponente der Interaktion zwischen Untersucher und Messzeitpunkt
Varianz des Residualfehlers	
σ^2_{PMR}	Varianzkomponente des Residualfehlers

M = Messzeitpunkt; P = Proband; R = Untersucher; RM = Interaktion Untersucher und Messzeitpunkt; PMR = Restfehler; PR = Interaktion Proband und Untersucher

Für die Bedingungen (2) G-inter-rater und (3) G-inter-Messzeitpunkt wurden jeweils 3 unterschiedliche Varianten mit fixen und zufälligen Facetten durchgespielt:

- (a) Messzeitpunkt und Untersucher sind zufällige Facetten
- (b) die G-Facette ist zufällig
- (c) die G-Facette ist fix

Für alle 3 Bedingungen wurden die universalen Werte und die absoluten sowie relativen Fehler abgeleitet (siehe Anhang 1–3).

Datenanalyse

G-Studie

Entsprechend der G-Theorie wurden für alle Haupteffekte, Interaktionseffekte und für den Residualfehler Varianzkomponenten berechnet (▶ **Tab. 5**). Für die Berechnung der Varianzkomponenten wurde jeweils der Durchschnitt der isometrischen Maximalkraft der Muskelgruppe Schulterabduktoren verwendet. Diese wurden mit einem linearen Mixed-Model (lmer) des R-lme4-package, einer Funktion des Softwareprogrammes R, Version 3.6.1 für Windows, geschätzt (R-Code siehe Anhang 7) [13, 14]. Aus den Varianzen wurden die Varianzen der universalen Werte (σ^2_T) und der absoluten (σ^2_Δ) sowie der relativen Fehler (σ^2_δ) gebildet (siehe Anhang 4). Abschließend wurden die G-Koeffizienten ($E\rho^2$) respektive (Φ) für alle Bedingungen empirisch berechnet (siehe Anhang 5). Zur Interpretation der G-Koeffizienten wurde folgendes Klassifikationssystem verwendet: Werte < 0,4 bedeutet schlechte, 0,4–0,75 faire bis gute und > 0,75 exzellente Reliabilität [11].

▶ **Tab. 6** Demografische und gesundheitsbezogene Daten der Probanden (n = 50).

Eigenschaften		Werte
Geschlecht, n (%)	Frauen	38 (76)
	Männer	12 (24)
Alter, in Jahren	Mittelwert ± SD	56 ± 14
	Range	21–82
BMI, in kg/cm²	Mittelwert ± SD	26 ± 5
Diagnose, n (%)	DM	22 (44)
	PM	17 (34)
	assoziiert	11 (22)
Krankheitsdauer (Monate)	Mittelwert ± SD	36 ± 18
	Range	0.25–240

DM = Dermatomyositis; n = Häufigkeit; PM = Polymyositis; SD = Standardabweichung

▶ **Tab. 7** Durchschnittliche Kraft der Schulterabduktoren für alle Messzeitpunkte M1–M3.

M1	M2	M3
Newton Mittelwert ± SD	Newton Mittelwert ± SD	Newton Mittelwert ± SD
92 ± 39	88 ± 35	89 ± 42

M = Messzeitpunkt; SD = Standardabweichung

Ergebnisse

Von 50 Probanden waren 22 von Dermatomyositis betroffen, 17 hatten Polymyositis und 11 litten an einer assoziierten Myositis (Myositis kombiniert mit einer anderen rheumatologischen oder onkologischen Erkrankung). ▶ **Tab. 6** beschreibt die demografischen und gesundheitsbezogenen Daten der Probanden.

Die Probanden erreichten bei der ersten Messung eine durchschnittliche Maximalkraft von 92 N ± 39 N SD, bei der zweiten Messung 88 N ± 35 N SD und bei der dritten Messung 89 N ± 42 N SD (▶ **Tab. 7**).

G-Studie

Die größte Varianz war auf die Probanden zurückzuführen (σ^2_P = 1154,9; 77 %) . Die Varianzen der Untersucher und der Messzeitpunkte waren gering (σ^2_R = 7,6; 0,5 % und σ^2_M = 8,9; 0,6 %). Die größte Interaktionsvarianz bestand zwischen den Untersuchern und den Probanden (σ^2_{PR} = 164,3; 11 %). Es gab keine Interaktionsvarianz zwischen den Untersuchern und den Messzeitpunkten (σ^2_{RM} = 0; 0 %). Die Varianz des Residualfehlers lag bei σ^2_{PMR} = 160,6; 11 % (▶ **Tab. 8**).

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

▶ **Tab. 8** Varianzkomponenten und Prozentangaben aller Haupteffekte, Interaktionseffekte und des Residuals der isometrischen Maximalkraft der Schulterabduktoren.

Varianzquellen	Varianz-komponenten	Prozent %
Proband, σ^2_P	1154,9	77,0
Untersucher, σ^2_R	7,6	0,5
Messzeitpunkt, σ^2_M	8,9	0,6
Proband × Untersucher, σ^2_{PR}	164,3	11,0
Untersucher × Messzeitpunkt, σ^2_{RM}	0	0
Proband × Untersucher × Messzeitpunkt (Residualfehler), σ^2_{PMR}	160,6	11,0

M = Messzeitpunkt, P = Proband, PMR = Restfehler, PR = Interaktion Proband und Untersucher, R = Untersucher, RM = Interaktion Untersucher und Messzeitpunkt

Die G-Koeffizienten ($E\rho^2$) und (Φ) lagen zwischen 0,77 und 0,89. Dieses Ergebnis entspricht einer exzellenten Reliabilität. Für die Bedingung (1) G-overall war der G-Koeffizient (Φ = 0,77), für die Bedingung (2) G-inter-rater ($E\rho^2$ = 0,78 und Φ = 0,78) und für die Bedingung (3) G-inter-Messzeitpunkt ($E\rho^2$ = 0,89 und Φ = 0,89). Die Ergebnisse resultierten unabhängig davon, ob die G-Facette als „zufällige" oder „fixe" Facette behandelt worden war.

Reliabilitätsstudie

Die Interraterreliabilität, erhoben mit der KTT, war ebenfalls exzellent (ICC2: 0,88) [12]. Die Formel zur Berechnung des ICC ist im Anhang 6 dargestellt. Die Ergebnisse der Reliabilitätsstudie sind detailliert beschrieben in [12]. ▶ **Tab. 9** zeigt die Ergebnisse der G-Studie und der Reliabilitätsstudie.

▶ **Tab. 9** G- und Φ-Koeffizienten für alle Bedingungen sowie Intraklassenkorrelationskoeffizient für die Muskelgruppe Schulterabduktoren.

Reliabilitätsstudie	
Intraklassenkorrelationskoeffizient (ICC)	0,88
G-Studie	
G-overall	
Bedingungen	Φ-Koeffizient
Messzeitpunkt und Untersucher „zufällig" (1a)	0,77[1]
G-inter-rater	
Bedingungen	Φ-Koeffizient
Messzeitpunkt und Untersucher „zufällig" (2a)	0,78
Messzeitpunkt „fix" und Untersucher „zufällig" (2b)	0,78
	G-Koeffizient ($E\rho^2$)
Messzeitpunkt „zufällig" und Untersucher „fix" (2c)	0,78
G-inter-Messzeitpunkt	
Bedingungen	Φ-Koeffizient
Messzeitpunkt und Untersucher „zufällig" (3a)	0,89
Untersucher „fix" und Messzeitpunkt „zufällig" (3b)	0,89
	G-Koeffizient ($E\rho^2$)
Untersucher „zufällig" und Messzeitpunkt „fix" (3c)	0,89

[1] Berechnungsbeispiel siehe Anhang 7.5.1

Diskussion

Das Ziel dieser Arbeit war, in die Grundelemente der G-Theorie einzuführen und anhand eines Beispiels aus der Physiotherapie eine G-Studie exemplarisch durchzuführen. Auf diese Weise war der Vergleich der beiden Ansätze (KTT und G-Theorie) hinsichtlich der Ergebnisse möglich.

Die Ergebnisse zeigen, dass die Reliabilität – unabhängig von fixen oder zufälligen Facetten – für alle Bedingungen exzellent ist. Die Varianzen sind hauptsächlich auf die Probanden und die Probanden-Untersucher-Interaktion zurückzuführen. Dieses Ergebnis deutet darauf hin, dass das handgehaltene Dynamometer hinsichtlich der Schulterabduktoren sehr gut zwischen Personen mit IM differenziert. Zudem ist eine Verallgemeinerung für verschiedene Messzeitpunkte und Untersucher möglich. Aufgrund der geringen Varianzen der Haupteffekte „Untersucher" und „Messzeitpunkt" ist die Reliabilität für alle Bedingungen gleich – unabhängig von fixen oder zufälligen Facetten. Die G-Koeffizienten für Bedingung (2) (G-inter-rater) ist aufgrund der größeren Interaktionsvarianz zwischen Untersucher und Probanden und dem dadurch größeren Fehleranteil geringer als diejenigen der Bedingung (3) G-inter-Messzeitpunkt. Mehrere Gründe für die sehr guten Werte der G-Koeffizienten und die geringe Variabilität der Haupteffekte sowie der Interaktionseffekte kommen infrage: ein optimal standardisiertes Messprotokoll, sehr gute Eichung durch die Therapeutinnen sowie eine akkurate Patienteninstruktion.

Studien, welche die Reliabilität ebenfalls mit der G-Theorie untersuchten, erzielten ähnliche Werte [15–17]. Zwei Studien ermittelten die Kraft der Knie-Extensoren bei gesunden Personen mittels eines handgehaltenen Dynamometers [15, 17]. Eine Untersuchung bestimmte Knieflexoren und -extensoren bei Patienten nach Knie- oder Hüft-Endoprothetik mittels eines fixierten Dynamometers [16]. Die Berechnung der G-Koeffizienten erfolgte jedoch nur in 2 Studien. Das Ergebnis war jeweils exzellent inter-rater (Φ: 0,92–0,99) [16], inter-tag (Φ: 0,95) und overall (Φ: 0,93) [17].

Die exzellente Interraterreliabilität war unabhängig davon, ob die KTT (ICC2 = 0,88) [12] oder die G-Theorie ($E\rho^2$ = 0,78; Φ = 0,78) zum Einsatz kam. Gründe für die ähnlichen Werte (ICC und G-Koeffizienten) könnten darin bestehen, dass zur Berechnung des ICC2 die Fehlerquelle „Untersucher" ebenfalls

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

Berücksichtigung fand. Dadurch ist ICC2 äquivalent zum G-Koeffizienten, dessen Berechnung ebenfalls mit einer einzelnen zusätzlichen Fehlerquelle (Untersucher) erfolgte [17].

In der vorliegenden Studie wurde neben der Fehlerquelle „Untersucher" als weitere Fehlerquelle der Messzeitpunkt berücksichtigt. Die einzelnen Fehlerkomponenten wurden bei der Berechnung der G-Koeffizienten summiert. Daraus kann ein größerer Fehleranteil resultieren. Ebenso können Unterschiede in den ICC-Werten und G-Koeffizienten entstehen. In Studien, in denen die Reliabilität ebenfalls anhand beider Ansätze (KTT bzw. G-Theorie) ermittelt und die Ergebnisse miteinander verglichen wurden, waren ebenfalls geringe Unterschiede der ICC-Werte gegenüber G-Koeffizienten feststellbar [17, 18]. Da Messungen im klinisch-therapeutischen Alltag unter nicht vollständig kontrollierbaren Bedingungen stattfinden, sind die Ergebnisse durch eine Vielzahl von Fehlerquellen beeinflusst. Daher ist es wichtig, diese Fehlerquellen zu untersuchen.

Das Ziel einer G-Studie besteht darin, die Komponenten eines Messprozederes zu ermitteln, die am stärksten zur Messvariabilität beitragen. Dieses Vorgehen ermöglicht Physiotherapeuten zu entscheiden, wie sie das Messinstrument wirksam und effizient in der Praxis einsetzen. Das Ergebnis einer G-Studie kann beispielsweise darauf hinweisen, dass die Variabilität hauptsächlich auf den Messzeitpunkt zurückzuführen ist. Je nach untersuchtem Kontext könnte das bedeuten, dass (1) Messungen immer zur selben Tageszeit durchgeführt werden sollten oder (2) das Messinstrument sich nicht für Verlaufsmessungen eignet. Im Gegensatz zur KTT bietet die G-Theorie weitere Informationen bezüglich der Generalisierbarkeit aller untersuchten Fehlerquellen im Rahmen des Messprozederes.

In der Physiotherapie liegen bislang wenige Studien vor, welche die Reliabilität von Messinstrumenten mit dem Ansatz der G-Theorie untersuchen. In den durchgeführten Studien wurde die Reliabilität abhängig vom Kontext für unterschiedliche Bedingungen (Test-retest, inter-rater, inter-tag, inter-trial oder overall) ermittelt. Mehrere Ursachen könnten zur bisher geringen Verbreitung der G-Theorie in der physiotherapeutischen Forschung infrage kommen. Neben der geringen Bekanntheit könnten auch die hohen statistischen und technischen Anforderungen limitierend sein. Zudem ist die Auswertung großer Datensätze sehr aufwendig. Es ist bisher auch kaum bekannt, dass G-Studien mit ordinal skalierten Daten durchführbar sind [31, 32].

Limitationen

Eine Limitation dieser Arbeit bestand in der Zuteilung der Probanden zu Therapeutin A oder B. Dadurch ergab sich ein nicht vollständig gekreuztes Design. Bei einem vollständig gekreuzten Design wäre es möglich gewesen, eine zusätzliche Fehlerquelle einzuschätzen: die Interaktion zwischen Probanden und Messzeitpunkt. Diese Situation hätte eine weitere Präzisierung der Fehlerquellen ermöglicht. Zusätzlich hätte der Messfehler (SEM) für jede Bedingung aus der Wurzel der absoluten respektive relativen Fehlervarianz berechnet werden können, um daraus weiter die kleinste erkennbare Veränderung (SDC) zu berechnen. Diese Berechnung ist vor allem für Verlaufsmessungen wichtig, um Veränderungen (vorher-nachher) zu quantifizieren. Aus exemplarischen Gründen wurde für die Bedingungen (2) und (3) eine Variante mit fixer G-Facette durchgespielt. Die klinische Forschung ermittelt jedoch häufig, ob eine Generalisierung für zufällig gewählte Untersucher, Messzeitpunkte oder für beide Komponenten möglich ist. Deshalb finden vor allem die Varianten (a) und (b) oder die Bedingung (1) (Varianten mit zufälligen Facetten) Verwendung.

Schlussfolgerungen

Die Ergebnisse der exemplarischen G-Studie deuten darauf hin, dass die G-Theorie gegenüber der KTT Vorteile hat. Erstens ist es möglich, für beliebig viele Fehlerquellen eines Messprozederes das Ausmaß der Variabilität zu ermitteln. Zweitens lässt sich die Generalisierbarkeit eines Messprozederes für beliebig viele Facetten ermitteln. Dadurch können Physiotherapeuten im klinischen Alltag Messungen wirkungsvoll und effizient einsetzen.

Mittels einer D-Studie wäre es möglich, die beste Messstrategie zu ermitteln. Dabei gilt es, zu untersuchen, mit welcher Kombination der Untersucherzahl und der Messgelegenheiten die jeweils beste Generalisierbarkeit möglich ist. Dadurch könnte es gelingen, das kosteneffizienteste und präziseste Messprozedere mit den geringsten Fehleranteilen zu ermitteln.

Interessenkonflikt

Die Autorinnen/Autoren geben an, dass kein Interessenkonflikt besteht.

Literatur

[1] Oesch P, Eberhardt R. Bewegungsapparat. 3., überarbeitete und ergänzte Aufl. Bern: Hogrefe; 2017

[2] Vet HCWd. Measurement in Medicine: A Practical Guide. Cambridge: University Press; 2011. doi:10.1017/CBO9780511996214

[3] Denegar CR, Ball DW. Assessing Reliability and Precision of Measurement: An Introduction to Intraclass Correlation and Standard Error of Measurement; 1993; 2: 35. doi:10.1123/jsr.2.1.35

[4] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86: 420–428

[5] Streiner DL. Health measurement scales: a practical guide to their development and use. 5th ed. Oxford: University Press; 2015

[6] Naizer G. Basic concepts in generalizability theory: a more powerful approach to evaluating reliability. Paper, präsented Anual Meeting Southwest Edu Research Asociation. Houston, TX; 1992

[7] Shavelson RL, Webb NM. Generalizability Theory: A Primer. Thousand Oaks, CA: Sage; 1991

[8] Brennan RL. Generalizability theory. New York: Springer; 2001

[9] de Bruin ED, Rozendal RH, Stussi E. Reliability of phase-velocity measurements of tibial bone. Phys Ther 1998; 78: 1166–1174. doi:10.1093/ptj/78.11.1166

[10] Heitman RJ, Kovaleski JE, Pugh SF. Application of generalizability theory in estimating the reliability of ankle-complex laxity measurement. J Athl Train 2009; 44: 48–52. doi:10.4085/1062-6050-44.1.48

[11] Fleiss JL. Design and Analysis of Clinical Experiments; Wiley; 2011

[12] Baschung Pfister P, de Bruin ED, Sterkele I et al. Manual muscle testing and hand-held dynamometry in people with inflammatory myopathy: An intra- and interrater reliability and validity study. PloS one 2018; 13: e0194531. doi:10.1371/journal.pone.0194531

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.

[13] Bates D, Mächler M, Bolker B et al. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 2015; 67: 48. doi:10.18637/jss.v067.i01

[14] R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna; 2019

[15] Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. Phys Ther 1993; 73: 386–395; discussion 396–401. doi:10.1093/ptj/73.6.386

[16] Gagnon D, Nadeau S, Gravel D et al. Reliability and validity of static knee strength measurements obtained with a chair-fixed dynamometer in subjects with hip or knee arthroplasty. Arch Phys Med Rehabil 2005; 86: 1998–2008. doi:10.1016/j.apmr.2005.04.013

[17] Anthony A, Gatti PWS, Nicholas MB et al. How to Optimize Measurement Protocols: An Example of Assessing Measurement Reliability Using Generalizability Theory. doi:10.3138/ptc-2018-0110

[18] Macintyre NJ, Bennett L, Bonnyman AM et al. Optimizing reliability of digital inclinometer and flexicurve ruler measures of spine curvatures in postmenopausal women with osteoporosis of the spine: an illustration of the use of generalizability theory. ISRN Rheumatol 2011: 571698. doi:10.5402/2011/571698

[19] Bauer CM, Rast FM, Ernst MJ et al. Concurrent validity and reliability of a novel wireless inertial measurement system to assess trunk movement. J Electromyogr Kinesiol 2015; 25: 782–790. doi:10.1016/j.jelekin.2015.06.001

[20] Bauer CM, Heimgartner M, Rast FM et al. Reliability of lumbar movement dysfunction tests for chronic low back pain patients. Man Ther 2016; 24: 81–84. doi:10.1016/j.math.2016.02.013

[21] Lafave M, Butterwick D. A generalizability theory study of athletic taping using the Technical Skill Assessment Instrument. doi:10.4085/1062-6050-49.2.22. Im Internet (Stand: 29.02.2020): www.cochranelibrary.com/central/doi/10.1002/central/CN-01115682/full

[22] Levac D, Missiuna C, Wishart L et al. The motor learning strategy instrument: interrater reliability within usual and virtual reality physical therapy interventions. Pediatr Phys Ther 2013; 25: 53–60. doi:10.1097/PEP.0b013e3182750c28

[23] Watkins B, Darrah J, Pain K. Reliability of passive ankle dorsiflexion measurements in children: comparison of universal and biplane goniometers. Im Internet (Stand: 29.02.2020): www.cochranelibrary.com/central/doi/10.1002/central/CN-00182435/full

[24] Ernst MJ, Sommer BB, Meichtry A et al. Intra-rater reliability of determining positions of cervical spinous processes and measuring their relative distances: An update to define rigid bodies of the cervical spine in a movement laboratory setting. BMC research notes 2019; 12: 265. doi:10.1186/s13104-019-4299-8

[25] Lariviere C, Ludvig D, Kearney R et al. Identification of intrinsic and reflexive contributions to low-back stiffness: medium-term reliability and construct validity. J Biomech 2015; 48: 254–261. doi:10.1016/j.jbiomech.2014.11.036

[26] Pryseley A, Ledent EY, Drewes AM et al. Applying concepts of generalizability theory on data from experimental pain studies to investigate reliability. Basic & clinical pharmacology & toxicology 2009; 105: 105–112. doi:10.1111/j.1742-7843.2009.00408.x

[27] Rast FM, Graf ES, Meichtry A et al. Between-day reliability of three-dimensional motion analysis of the trunk: A comparison of marker based protocols. J Biomech 2016; 49: 807–811. doi:10.1016/j.jbiomech.2016.02.030

[28] Swaine BR, Lortie E, Gravel D. The reliability of the time to execute various forms of the finger-to-nose test in healthy subjects. Physiotherapy theory and practice 2005; 21: 271–279. doi:10.1080/09593980500321119

[29] Fortin C, Feldman DE, Cheriet F et al. Reliability of a quantitative clinical posture assessment tool among persons with idiopathic scoliosis. Physiotherapy 2012; 98: 64–75. doi:10.1016/j.physio.2010.12.006

[30] Sommer BB, Weisenhorn M, Ernst MJ et al. Concurrent validity and reliability of a mobile tracking technology to measure angular and linear movements of the neck. J Biomech 2019; 96: 109340. doi:10.1016/j.jbiomech.2019.109340

[31] Vangeneugden T, Laenen A, Geys H et al. Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. Biometrics 2005; 61: 295–304. doi:10.1111/j.0006-341X.2005.031040.x

[32] Vispoel WP, Morris CA, Kilinc M. Applications of generalizability theory and their relations to classical test theory and structural equation modeling. Psychol Methods 2018; 23: 1–26. doi:10.1037/met0000107

Dieses Dokument wurde zum persönlichen Gebrauch heruntergeladen. Vervielfältigung nur mit Zustimmung des Verlages.