Braz J Phys Ther 29 (2025) 101193

Contents lists available at ScienceDirect

Brazilian Journal of Physical Therapy

journal homepage: www.elsevier.com/locate/bjpt

Masterclass

Considerations when designing, analyzing, and reporting reliability studies

Paul Stratford, Gregory F. Spadoni, Ayse Kuspinar, Luciana Macedo *

School of Rehabilitation Science, Faculty of Health Science, McMaster University, Hamilton, Ontario, Canada

ARTICLE INFO

Keywords:
Measurement
Methods
Rehabilitation
Reliability
Research design
Statistics

ABSTRACT

Background: Reliability studies have a long history in the physical therapy literature and their sophistication has evolved over the decades. Often, however, there has been incomplete reporting or a lack of coherence among study purpose, design, choice of analysis, sample size justification, and reporting of results. Two possible explanations for this oversight are a vaguely written purpose statement and statistical software that does not provide all essential information.
Objective: The goal of this masterclass is to provide considerations and resources to assist investigators structure a coherent reliability study design and subsequent presentation of results.
Discussion: This masterclass highlights the importance of framing a study purpose that clearly distinguishes between a hypothesis testing and parameter estimation study and appropriately labelling the study design. It also stresses the importance of stating whether the raters are the only ones of interest or whether they are intended to represent a larger group of raters, applying a sample size calculation consistent with the study purpose, and reporting results that align with the study purpose and design.

Introduction

Reliability studies have had a long history in the physical therapy literature.[1,2] These studies have primarily addressed inter-rater, test-retest, or a combination of the two designs. Early publications presented point estimates of intraclass correlation coefficients (ICCs)[3] and later reports accompanied the ICCs with 95% confidence intervals (CIs).[4] Near the end of the 20th century, the standard error of measurement (SEM) appeared with increased regularity,[5] and by 2010 it was commonplace to find ICCs, their 95% CIs, and point estimates of SEMs included in many reliability studies of interest to physical therapists. Although there is a rich history of reliability studies in the physical therapy literature, often these studies have shown a lack of coherence among the purpose, design, appropriate choice of analysis, reporting, and interpretation of results. For example, reliability studies (1) rarely include a sample size calculation or exhibit coherence among the purpose statement, sample size calculation, and analysis;[6–8] (2) occasionally apply an inappropriate ICC form that ignores a meaningful difference in mean scores between test and retest;[8–14] and (3) have reported SEMs without including CIs even though ICCs are reported with CIs.[9–13,15] We believe that vaguely written research questions and statistical software packages that do not provide all relevant analyses play important roles in contributing to the lack of coherence often seen in reliability studies. Better reporting standards including those advocated by the Consensus-based Standards for the Selection of Health Measurement Instruments group and Guidelines for Reporting Reliability and Agreement Studies (GRRAS) are required.[16,17] The goal of this monograph is to provide considerations and resources to assist investigators structure a coherent reliability study design and subsequent presentation of results.

Before proceeding there are two essential points to acknowledge. First, reliability is not a property of a measure, but rather of a measure's scores or measured values.[18,19] Messick states, "Tests do not have reliabilities and validities, only test responses do. This is an important point because test responses are a function not only of the items, tasks, or stimulus conditions but of the persons responding and the context of measurement."[18] The second point is that reliability is not an all-or-none property: it exists to a degree. Deciding about the adequacy of reliability in a specific context requires evaluating the extent to which measured values differentiate among the objects of measure (hereafter referred to as patients) and the absolute error expressed in the same unit as the original measurement. As pointed out by the GRASS group, tagging adjectives (e.g., poor, fair, moderate, substantial, almost perfect) to ICC values is not enough.[17] Making a judgment based on both relative (ICC) and absolute (SEM) reliability coefficients is necessary. Having a firm understanding of these fundamentals will discourage investigators from

* Corresponding author at: 1400 Main St. W. Room 441, IAHS, Hamilton, ON, L8S 1C7.
E-mail address: macedol@mcmaster.ca (L. Macedo).

https://doi.org/10.1016/j.bjpt.2025.101193
Received 2 July 2024; Received in revised form 2 October 2024; Accepted 14 February 2025
Available online 28 February 2025

presenting the often-seen concluding statement that declares a measure reliable, and the subsequent influence this authoritative pronouncement may have on readers.[20,21]

Framing the study purpose

A clearly crafted purpose statement conveys the context and sets the stage for what is to follow in the design, sample size calculation, and analysis (Fig. 1). Too often vague verbs such as "to investigate", "to examine", and "to explore" are found in purpose statements.[9,10,22] These verbs lack the specificity needed to allow the seamless transition from purpose statement to research question. Better verbs such as "to determine" and "to estimate", for example, direct a reader's expectation towards hypothesis testing, where the study's obtained reliability coefficient will be formally compared to the hypothesized null value, or parameter estimation, where point and interval estimates (e.g., 95 % CI) for the likely location of the population's reliability value will be reported. If the goal is hypothesis testing, including null and alternate hypotheses further clarify the investigator's intent concerning the directionality of the subsequent statistical test (i.e., 1- or 2-tailed). Essential components of parameter estimation purpose statements include specification of the confidence level of interest (e.g., 95 % CI) and whether the interval of interest is 1- or 2-sided.

Hypothesis testing example

Purpose statement: The purpose of this study was to determine if the inter-rater reliability, as quantified by a Form 2,1 Shrout and Fleiss ICC (hereafter identified as $ICC_{2,1}$),[23] of the Chedoke Arm and Hand Activity Inventory (CAHAI)[24] in patients post-stroke fulfilling the eligibility criteria exceeds 0.90.

Research Question: Does the inter-rater reliability, as quantified by a Shrout and Fleiss $ICC_{2,1}$,[23] of CAHAI[24] scores from patients post-stroke fulfilling the eligibility criteria exceed 0.90? The answer to this clearly stated question will be either "Yes" or "No" as determined by the critical p-value (e.g., $p < 0.05$).

Null hypothesis: CAHAI test scores from patients post-stroke fulfilling the eligibility criteria will not demonstrate sufficient reliability for clinical application ($ICC_{2,1} \leq 0.90$).[25,26]

Alternate hypothesis: CAHAI test scores from patients post-stroke fulfilling the eligibility criteria will demonstrate sufficient reliability for clinical application ($ICC_{2,1} > 0.90$).[25,26]

Parameter estimation example

Purpose statement: The purpose of this study was to estimate the inter-rater reliability of CAHAI scores as quantified by Shrout and Fleiss $ICC_{2,1}$ and 2-sided 95% CI, in patients post-stroke fulfilling the eligibility criteria.

Research question: To what extent are rater assigned CAHAI scores reliable, as quantified by a Shrout and Fleiss $ICC_{2,1}$ and 2-sided 95% CI, in patients post-stroke fulfilling the eligibility criteria? The answer to this question will be point and interval estimates of the ICC.

Design

Typical designs appearing in the physical therapy literature include intra-rater, inter-rater, test-retest, and a combination of rater and test-retest designs. An assumption for all reliability designs is that the feature being measured does not change during the course of the study: a patient's true score does not change. In this commentary we restrict our discussion to the frequently seen design where a single rating is obtained by either each rater in an inter-rater reliability study, or a single rating at each occasion in an intra-rater or test-retest reliability study.

Sample variability impacts the magnitude of an ICC. All else being equal, a sample with a larger variability—think a wider range of

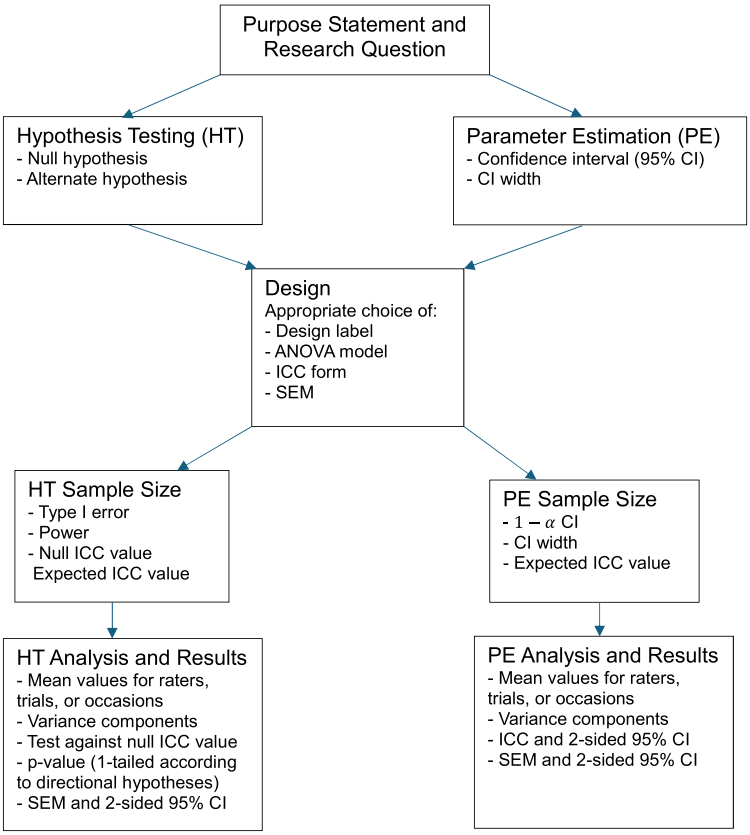

Fig. 1. Design and analysis components of a reliability study.

values—will have a higher ICC than a sample with less variability. For this reason it is critical that the sample is representative of the population of interest. This is an important consideration at two levels. First, it is important that the study patients are representative of those to whom the results will be applied. An ideal sampling strategy for reliability studies would be random sampling from a larger pool of individuals fulfilling the eligibility criteria. Often, however, this is not possible because investigators do not have access to a sufficient number of patients at a single point in time. When this is the case, consecutive sampling of patients fulfilling the eligibility criteria is a reasonable method for obtaining an unbiased sample.

When raters are involved in the measurement process, knowledge of their characteristics and who they are intended to represent is necessary. In addition to mentioning the raters' experiences with patients similar to the study sample, and whether the raters have had training on the outcome measure, it is essential for an investigator to state whether the raters taking part in the reliability study are viewed as the only raters of interest, or whether they are intended to represent a larger group of raters. We will subsequently see in the analysis section that this distinction has an important implication for the choice of analysis of variance (ANOVA) model from which the ICC and SEM are obtained.

An example of a pure inter-rater reliability study design would be clinicians observing and scoring the same stimulus (e.g., CAHAI scores assigned to videotaped performances of a group of patients). An example, of a pure test-retest reliability study would be patients completing a self-report functional status measurement over an interval when no true change is believed to have occurred. Often however, owing to feasibility and the importance of providing a real-world context to the assessment process (i.e., video-taped ratings don't account for differences eliciting test results), a design that combines both inter-rater (or intra-rater) and test-retest reliability components (a patient provides separate performances or stimuli for each rater or occasion) is necessary. For example, a design where four raters independently administer and score the CAHAI would include variations between raters and patient's performances. Not only does an interpretive challenge arise when an investigator assigns the intra-rater, inter-rater, or test-retest label to the combined design, but the combined design has also led some investigators to apply a Form 3 ICC, for example, when a Form 2 ICC is more appropriate.[8–13] We will elaborate on this peculiarity in the analysis section. For a combined design it may be more informative to comment on sources of variation (see Appendix for variance calculation) rather than assigning a procrustean label to the study design. For example, an investigator could state the following: "Our design has components of inter-rater and test-retest reliability. Accordingly, apparent differences among raters' scores include both differences among raters and inherent variation in patients' performances."

Another important consideration in an inter-rater reliability study—either pure or combined design—is the potential effect the order of testing may have on the results. If the order of testing among raters were the same for all patients, it would be impossible to determine whether a systematic difference among raters was owing to a rater or the order of testing associated with that rater. Although some investigators have randomized the order of testing to raters, randomization alone does not ensure balance, particularly in reliability studies with relatively small sample sizes. The Latin Square design provides one remedy for addressing this concern.[27] In a Latin Square design the order of testing is perfectly balanced among raters; an example is shown in Fig. 2. Notice that in this example each rater's assessment order precedes any other rater's order the same number of times.

Sample size estimation

Hypothesis testing sample size

Several sample size estimation methods exist for hypothesis testing reliability studies,[25,26,28] one of which is illustrated below.[28] Before proceeding it is important to acknowledge that the distribution of reliability coefficients is non-normal. Accordingly, the first step in a sample size calculation is to transform the expected and null reliability coefficient values to a distribution that approximates a normal distribution. This is accomplished with Fisher's Z-transformation.[29] The expected reliability (R_E) is what an investigator anticipates finding in the study, and the null reliability value (R_0) is that specified in the null hypothesis statement.

Fisher's Z-Transformation of expected reliability value R_E

$$Z_E = 0.5\ln\left(\frac{1+(k-1)R_E}{1-R_E}\right)$$

Fisher's Z-Transformation of null reliability value R_0

$$Z_0 = 0.5\ln\left(\frac{1+(k-1)R_0}{1-R_0}\right)$$

where ln is the natural logarithm and k is the number of repeated measurements (raters or occasions)

The transformed Z-values are then applied to the following sample size formula which is based on a 1-way ANOVA model.[28] Estimates from this model will be conservative when a two-way model is applied.

$$n = \left(\frac{k\ (Z_\alpha + Z_\beta)^2}{2\ (Z_E - Z_0)^2\ (k-1)}\right)$$

where k, Z_E, and Z_0 have been defined previously, and $Z_\alpha$and Z_β represent the standard normal deviates for Type I and Type II errors respectively. For a Type I error of 0.05 and a Type II error of 0.20 these Z-values would be 1.64 (1-tailed owing to the directional hypotheses) and 0.84, respectively.

Parameter estimation sample size 2-sided confidence interval

Often investigators are interested in gaining an impression of a reasonable range of values in which the population reliability coefficient is likely to lie. When this is the case a 2-sided CI is desirable and can be estimated as follows[30]

$$n = \frac{\left(8\ Z^2_{\alpha/2}\ \left[(1-R_E)^2\ (1+(k-1)R_E)^2\right]\right)}{(k\ (k-1)\ w^2)}\ +\ 1$$

where $Z_{\alpha/2}$ is 1.96 for a 2-sided 95% CI; k is number of raters, occasions,

	Order of Testing			
	1st	2nd	3rd	4th
Patient 1	Rater 1	Rater 2	Rater 3	Rater 4
Patient 2	Rater 2	Rater 1	Rater 4	Rater 3
Patient 3	Rater 3	Rater 4	Rater 1	Rater 2
Patient 4	Rater 4	Rater 3	Rater 2	Rater 1

Fig. 2. Latin Square balanced for four raters.

or trials; w is the CI width; R_E is the expected ICC value.

Analysis

Given the importance of context specificity, it is essential to report patients', and when applicable, raters' descriptive characteristics.[16,17] Also, summary statistics describing rater or occasion mean values is necessary as they communicate in familiar units the extent to which a systematic difference among raters or between occasions exists.

For a test or measure to be clinically useful, it must have a sufficiently high ICC and a sufficiently low SEM. What constitutes "sufficient" will be context specific. Accordingly, reporting parameter estimates, or hypothesis test results of both the ICC and SEM are essential. Information necessary to calculate ICCs and SEMs is obtained from ANOVA tables. Form 1 ICCs and their corresponding SEMs are calculated from a 1-way ANOVA model, and Forms 2 and 3 ICCs and their related SEMs are calculated from a 2-way ANOVA model.

Intraclass correlation coefficients

Many popular statistical software packages provide hypothesis testing and CI options for ICCs. This is fortunate because the calculations required to determine the appropriate degrees of freedom are extensive and detailed by Shrout and Fleiss.[23] In addition to the formulae provided by Shrout and Fleiss, CIs can also be estimated using a bootstrap procedure where sampling with replacement is applied *and is appropriate when the statistical distribution is unknown or the assumption of normality is not satisfied*.[31]

Although there are many forms of ICCs,[32,33] we will restrict our discussion to those presented by Shrout and Fleiss where each rater (*or each occasion in a test-retest design: our words*) provides only a single measured value per patient.[23] In their seminal article, Shrout and Fleiss introduced six forms of ICCs in the context of an inter-rater reliability study (Table 1).[23] To assist investigators in choosing among these ICC forms, Shrout and Fleiss posed the following three questions: (1) "Is a one-way or two-way ANOVA model appropriate for the analysis of the reliability study?" (2) "Are differences between judges' *(or occasions: our words)* mean ratings relevant to the reliability of interest?" (3) "Is the unit of analysis an individual rating or the mean of several ratings?"[23]

A Form 1 ICC is appropriate when there is no natural structure linking repeated measurements and it is rarely misapplied. For example, different patients are assessed by different combinations of raters. However, making the appropriate choice between a Form 2 and Form 3 ICC has been challenging for some investigators, particularly when the design combines intra-rater or inter-rater, and test-retest components.[8–13] We believe the source of the problem lies in the investigators' applications of a pure rater design (i.e., Shrout and Fleiss' illustration) to that of a combined design. To better understand the source of the problem it is informative to compare the Forms 2 and 3 ICCs shown in Table 1. The salient feature is that although both forms are based on a 2-way ANOVA model, the denominator of the Form 2 ICC calculation includes the variance component associated with a systematic difference among repeated measurements, whereas the Form 3 ICC excludes this variance component. Therefore, when a systematic difference among repeated measurements exists, the Form 2 ICC will be less than the Form 3 ICC. We will subsequently see that this also impacts the SEM calculation. We provide three examples to provide clarity when choosing between Forms 2 and 3 ICCs.

Example 1. When a reliability study has a test-retest component where patients' performances are obtained for different trials or occasions, it is important to know whether a systematic difference exists. Accordingly, a Form 2 ICC is appropriate regardless of whether the same rater plays a meaningful role in obtaining the measured value or not.

Example 2. A Form 2 ICC is appropriate when reporting a pure inter-rater (or intra-rater) design where the raters are intended to represent a larger group of raters.

For example, an investigator is interested in estimating the inter-rater reliability and determining whether a systematic difference among raters is likely following a standardized training program for the CAHAI. The investigator is interested in generalizing the results to all raters who take part in future standardization training programs. Thus, knowing the extent to which a systematic difference is likely to occur in the population is important not only for the current investigator, but also to a wider audience.

Example 3. A Form 3 ICC is appropriate when reporting a pure inter-rater (or intra-rater) design where the reliability study raters are the only raters of interest.

A team of investigators will be undertaking a randomized clinical trial to evaluate two rehabilitation programs for patients post-stroke. The CAHAI will be the primary outcome measure, and it will be administered by four raters who have been hired for this project. In the clinical trial, each patient will be assessed pre- and post-intervention by one rater. Although it would be ideal to have both assessments performed by the same rater, the investigators believe it is possible that for

Table 1
ANOVA models, ICC forms, and SEM calculations.

ICC Form	Description	ICC Forms Obtained From Variance Components*	SEM Calculation
1,1 1, k	1-way ANOVA: **Source SS df MS** Patients SSP df_p MSP Within SSW df_w MSW Patients need not have the same number of measured values, nor do the measurements need to be performed by the same rater or set of raters	$ICC_{1,1} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_w^2}$ $ICC_{1,k} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_w^2}{k}}$	$SEM_1 = \sqrt{MSW} = \sqrt{\sigma_w^2}$ $SEM_k = \sqrt{\sigma_w^2/k}$
2,1 2, k	2-way ANOVA without an interaction term: **Source SS df MS** Patients SSP df_p MSP Rater/ or Occasion SSO df_o MSO Error SSE df_e MSE All patients are rated by the same set of raters which are considered to represent a larger group of raters	$ICC_{2,1} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_o^2 + \sigma_e^2}$ $ICC_{2,k} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_o^2}{k} + \frac{\sigma_e^2}{k}}$	$SEM_1 = \sqrt{\sigma_o^2 + \sigma_e^2}$ $SEM_k = \sqrt{(\sigma_o^2 + \sigma_e^2)/k}$
3,1 3, k	2-way ANOVA without an interaction term: **Source SS df MS** Patients SSP df_p MSP Rater or Occasion SSO df_o MSO Error SSE df_e MSE All patients are rated by the same set of raters which are the only raters of interest	$ICC_{3,1} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2} ICC_{3,k} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_e^2}{k}}$	$SEM_1 = \sqrt{MSE} = \sqrt{\sigma_e^2}$ $SEM_k = \sqrt{\sigma_e^2/k}$

Terms mentioned in the table.
SS sum or squares.
df degrees of freedom.
MS mean square.
σ_p^2 is the patient variance.
σ_w^2 is the within patient variance.
σ_o^2 is the rater or occasion variance.
σ_e^2 is the error variance.
k is the number of measurements averaged.
ICC intraclass correlation coefficient.
SEM standard error of measurement.

some patients different raters will perform the pre- and post-assessments in the subsequent clinical trial. In the reliability study each patient is assessed by all four raters. Because the reliability study raters are also the same and only raters performing the assessments in the subsequent clinical trial, a Form 3 ICC is appropriate. The reasoning is that if a systematic difference is identified in the reliability study, for example, *Rater 3*'s mean rating is 4-points more than the other raters, the investigator can correct for this bias for all of Rater 3's rating by subtracting 4-points. Accordingly, the systematic rater variance in the reliability coefficient calculation is effectively reset to zero and can be removed from the ICC calculation. The Form 3 ICC is appropriate.

Standard errors of measurement

Although it is now customary to include point estimates of the SEM in reliability studies, they are rarely accompanied by a CI. A likely explanation is that many statistical software packages do not provide these results. However, estimates of the SEM and 95% CI can be obtained by applying the information shown in Table 1 and the following formula[5,34]:

$$\frac{SSE}{\chi^2_{1-\alpha/2,\ df_e}};\ \frac{SSE}{\chi^2_{\alpha/2,\ df_e}}$$

where SSE is the sum of squares within or error from a 1-way or 2-way ANOVA respectively. $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are the chi-square values associated with the lower and upper confidence limits of interest, and df_e is the degrees of freedom for the error term from the appropriate ANOVA model.

Summary

The goal of this monograph was to provide considerations and resources to assist investigators design reliability studies and report their results in the physical therapy literature. This work is based on existing guidelines[16,17] and our review of published reliability studies, where frequently seen limitations included vaguely written purpose statements, mislabeling of study designs, not reporting whether raters were intended to represent a larger group of raters or the only raters of interest, failure to provide appropriate sample size calculations, application of inappropriate ICC forms, and not reporting CIs for the SEMs. In closing we propose that the most important consideration is to provide a clear and specific purpose statement and to ensure that each step of the design and subsequent reporting is true to this statement.

Declaration of competing interest

None.

Appendix

Variance Components and ICCs

In this example four raters independently administered and scored the CAHAI on 32 patients fulfilling the study's eligibility criteria. The order of testing was balanced among raters. Accordingly, apparent differences among raters also includes inherent differences within a patient's performance across the four occasions.

Summary Statistics by Rater

Rater 1 Mean, SD, N	Rater 2 Mean, SD, N	Rater 3 Mean, SD, N	Rater 4 Mean, SD, N
37.1, 25.1, 32	38.4, 24.2, 32	41.3, 23.8, 32	36.8, 24.2, 32

ANOVA and Variance Calculations

Source	Sum of Squares	DF	Mean Square	Variance Calculation	Variance Components
Patient	71,572.93	31	2308.80	$\frac{MSP - MSE}{N_{raters}} = \frac{2308.80 - 19.29}{4}$	572.38
Rater/Occasion	389.46	3	129.82	$\frac{MSRO - MSE}{N_{patients}} = \frac{129.82 - 19.29}{32}$	3.45
Error	1794.29	93	19.29	MSE = 19.29	19.29

$$ICC_{2,1} = \frac{572.38}{572.38 + 3.45 + 19.29}\ ICC_{3,1} = \frac{572.38}{572.38 + 19.29}$$

$$ICC_{2,1} = 0.96\ ICC_{3,1} = 0.97$$

$$SEM = 4.77\ SEM = 4.39$$

References

1. Rothstein JM. *Measurement in Physical Therapy*. New York: Churchill Livingstone; 1985.
2. Lamb R, Bohannon R, Craik RL, et al. Reliability discussion required. *Phys Ther*. 1987;67(4):501. https://doi.org/10.1093/ptj/67.4.501.
3. Krebs DE. Declare your ICC type. *Phys Ther*. 1986;66(9):1431. https://doi.org/10.1093/ptj/66.9.1431.
4. Stratford P. Confidence limits for your ICC. *Phys Ther*. 1989;69(3):237–238. https://doi.org/10.1093/ptj/69.3.237.
5. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther*. 1997;77(7):745–750. https://doi.org/10.1093/ptj/77.7.745.
6. Kwan L, Nishihara K, Cheung A, et al. Reliability and feasibility of extended field of view ultrasound imaging techniques for measuring the limb muscle cross-sectional area. *Physiother Can*. 2020;72(2):149–157. https://doi.org/10.3138/ptc-2018-0105.

7. Fonteyne L, Guinois-Cote S, Perugino L, et al. Interrater reliability among novice raters in the assessment of pelvic floor muscle tone using the reissing tone scale. *Physiother Can.* 2021;73(4):313–321. https://doi.org/10.3138/ptc-2019-0093.
8. Steffen T, Seney M. Test-retest reliability and minimal detectable change on balance and ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease rating scale in people with parkinsonism. *Phys Ther.* 2008;88(6): 733–746. https://doi.org/10.2522/ptj.20070214.
9. Dal Bello-Haas V, Klassen L, Sheppard MS, Metcalfe A. Psychometric properties of activity, self-efficacy, and quality-of-life measures in individuals with parkinson disease. *Physiother Can.* 2011;63(1):47–57. https://doi.org/10.3138/ptc.2009-08.
10. Bruyneel AV, Reinmann A, Sordet C, et al. Reliability and validity of the trunk position sense and modified functional reach tests in individuals after stroke. *Physiother Theory Pract.* 2024;40(1):118–127. https://doi.org/10.1080/09593985.2022.2101407.
11. Molhemi F, Monjezi S, Mehravar M, Shaterzadeh-Yazdi MJ, Majdinasab N. Validity, reliability, and responsiveness of Persian version of mini-balance evaluation system test among ambulatory people with multiple sclerosis. *Physiother Theory Pract.* 2024; 40(3):565–575. https://doi.org/10.1080/09593985.2022.2119908.
12. Cavalheiro Puzzi V, Mara Oliveira J, Bessa Alves T, et al. Validity and reliability of the Glittre-ADL test in adults with asthma. *Physiother Theory Pract.* 2023;39(5): 1052–1060. https://doi.org/10.1080/09593985.2022.2114301.
13. Wang S, Mani R, Zeng J, Chapple CM, Ribeiro DC. Test-retest reliability of movement-evoked pain and sensitivity to movement-evoked pain in patients with rotator cuff-related shoulder pain. *Braz J Phys Ther.* 2023;27(4), 100535. https://doi.org/10.1016/j.bjpt.2023.100535.
14. Calixtre LB, Fonseca CL, Gruninger B, Kamonseki DH. Psychometric properties of the Brazilian version of the Bournemouth questionnaire for low back pain: validity and reliability. *Braz J Phys Ther.* 2021;25(1):70–77. https://doi.org/10.1016/j.bjpt.2020.02.003.
15. Antunes AAM, Furtado SRC, Magalhães LC, Kirkwood RN, Vaz DV. Brazilian versions of the measure of processes of care-20 and measure of processes of care-service providers: translation, cross-cultural adaptation and reliability. *Braz J Phys Ther.* 2020;24(2):144–151. https://doi.org/10.1016/j.bjpt.2019.02.013 [published Online First: 20190227].
16. Mokkink LB, Terwee CB, Gibbons E, et al. Inter-rater agreement and reliability of the COSMIN (Consensus-based Standards for the selection of health status measurement instruments) checklist. *BMC Med Res Methodol.* 2010;10:82. https://doi.org/10.1186/1471-2288-10-82.
17. Kottner J, Audige L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol.* 2011;64(1):96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002.
18. Messick S. Validity. In: Linn RL, ed. *Educational Measurement.* 3rd ed. Phoenix: ORYZ Press; 1993:14.
19. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use.* 4th ed. New York: Oxford University Press; 2008.
20. Thoomes E, Cleland JA, Falla D, Bier J, de Graaf M. Reliability, measurement error, responsiveness, and minimal important change of the patient-specific functional scale 2.0 for patients with nonspecific neck pain. *Phys Ther.* 2024;104(1). https://doi.org/10.1093/ptj/pzad113.
21. Ansanello W, Dos Reis FJJ, Tozzo MC, et al. Reliability and validity of the avoidance of daily activities photo scale for patients with shoulder pain (adap shoulder scale). *Phys Ther.* 2023;103(12). https://doi.org/10.1093/ptj/pzad101.
22. Liz L, da Silva TG, Michaelsen SM. Validity, reliability, and measurement error of the remote fugl-meyer assessment by videoconferencing: tele-FMA. *Phys Ther.* 2023; 103(8). https://doi.org/10.1093/ptj/pzad054.
23. Shrout PE, Fleiss JL. Intraclass correlation: uses in assessing rater reliability. *Psychol Bull.* 1979;86(2):420–428.
24. Barreca S, Gowland CK, Stratford P, et al. Development of the chedoke arm and hand activity inventory: theoretical constructs, item generation, and selection. *Top Stroke Rehabil.* 2004;11(4):31–42. https://doi.org/10.1310/JU8P-UVK6-68VW-CF3W.
25. Donner A, Eliasziw M. Sample size requirements for reliability studies. *Stat Med.* 1987;6(4):441–448. https://doi.org/10.1002/sim.4780060404.
26. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med.* 1998;17(1):101–110. https://doi.org/10.1002/(sici)1097-0258(19980115)17:1.
27. Fleiss JL. *The Design and Analysis of Clinical Experiments.* New York: John Wiley & Sons; 1986.
28. Stratford PW, Spadoni GF. Sample size estimation for the comparison of competing measures' reliability coefficients. *Physiother Can.* 2003;55:225–229.
29. Fisher RA. *Statistical Methods for Research Workers.* 14th ed. Darien, CT: Hafner Publishing Company; 1970.
30. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. *Stat Med.* 2002;21(9):1331–1335. https://doi.org/10.1002/sim.1108.
31. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am Statistician.* 1983;37:36–48.
32. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods.* 1996;1(1):30–46. https://doi.org/10.1037//1082-989X.1.4.390.
33. Eliasziw M, Young SL, Woodbury MG. Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Phys Ther.* 1994;74(8):777–788. https://doi.org/10.1093/ptj/74.8.777.
34. Armitage P, Berry G. *Statistical Methods in Medical Research.* 3rd ed. New York: John Wiley and Sons; 1994.