ARTICLE

How to Optimize Measurement Protocols: An Example of Assessing Measurement Reliability Using Generalizability Theory

Anthony A. Gatti, MSc;[*] *Paul W. Stratford, MSc, PT;*[*] *Nicholas M. Brisson, PhD;*[*†] *Monica R. Maly, PhD, PT*[*‡]

ABSTRACT

Purpose: This article identifies how to assess multiple sources of measurement error and identify optimal measurement strategies for obtaining clinical outcomes. ***Method:*** Obtaining, interpreting, and using information gained from measurements is instrumental in physiotherapy. To be useful, measurements must have a sufficiently small measurement error. Traditional expressions of reliability include relative reliability in the form of an intra-class correlation coefficient and absolute reliability in the form of the standard error of measurement. Traditional metrics are limited to assessing one source of error; however, real-world measurements consist of many sources of error. The measurement framework generalizability theory (GT) allows researchers to partition measurement errors into multiple sources. GT further allows them to calculate the relative and absolute reliability of any measurement strategy, thereby allowing them to identify the optimal strategy. We provide a brief comparison of classical test theory and GT, followed by an overview of the terminology and methodology used in GT, and then an example showing how GT can be used to minimize error associated with measuring knee extension power. ***Conclusion:*** The methodology described provides tools for researchers and clinicians that enable detailed interpretation and understanding of the error associated with their measurements.

Key Words: biostatistics; outcome and process assessment, health care; rehabilitation research; reproducibility of results.

RÉSUMÉ

Objectif : décrire comment évaluer de multiples sources d'erreur de mesure et les stratégies de mesures optimales pour obtenir des résultats cliniques. ***Méthodologie :*** il est important d'obtenir, d'interpréter et d'utiliser l'information tirée des mesures en physiothérapie. Pour que ces mesures soient utiles, leur écart-type doit être suffisamment petit. Les expressions habituelles de fiabilité incluent la fiabilité relative sous forme de coefficient de corrélation intraclasse et la fiabilité absolue sous forme d'écart-type des mesures. Les mesures habituelles sont limitées à l'évaluation d'une source d'erreur. Cependant, les mesures réelles s'associent à plusieurs sources d'erreur. La théorie de généralisabilité (TG) du cadre de mesure permet aux chercheurs de diviser les erreurs de mesure selon de multiples sources. Elle leur permet également de calculer la fiabilité relative et absolue de toute stratégie de mesure, pour parvenir à une stratégie optimale. Le présent article fournit une brève comparaison entre la théorie du test classique et la TG, puis un aperçu de la terminologie et de la méthodologie utilisées en TG. Enfin, les auteurs présentent un exemple démontrant comment utiliser la TG pour limiter l'erreur associée à la mesure de la puissance d'extension du genou. ***Conclusion :*** la méthodologie décrite fournit des outils pour les chercheurs et les cliniciens afin de parvenir à une interprétation et une compréhension détaillées des erreurs de mesure.

Performing measurements, interpreting their results, and applying the information gained to shape clinical decisions are the cornerstones of physiotherapy practice. Having confidence in a measurement is essential in making decisions about a patient's status and determining whether a change has occurred. The smaller the measurement error, the greater the confidence in a measured value. Reliability is the measurement property that quantifies the confidence in a measured value. Like validity, reliability is context specific. It is not binary. Stated another way, tests and measures are not reliable; rather, a test's performance in a defined context exhibits a reliability that is expressed on a continuum. In this article, we briefly review classical test theory (CTT), introduce a

From the: [*]School of Rehabilitation Sciences, McMaster University, Hamilton; [‡]Department of Kinesiology, University of Waterloo, Waterloo, Ont.; [†]Julius Wolff Institute, Charité – Universitätsmedizin Berlin, Germany.

Correspondence to: Anthony A. Gatti, School of Rehabilitation Sciences, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4L8; gattia@mcmaster.ca.

Contributors: All authors designed the study; or collected, analyzed, or interpreted the data; and drafted or critically revised the article and approved the final draft.

Competing Interests: This research was supported by funding from Ontario Graduate Scholarships (AAG, NMB), the Arthritis Society Doctoral Salary Award (AAG), the Joint Motion Program (NMB), an operating grant (MRM No. 102643) from the Canadian Institutes of Health Research (CIHR), and a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (MRM No. 353715). The equipment required to carry out this research was purchased with funds received from the Canadian Foundation for Innovation Leaders Opportunity Fund and the Ministry of Research and Innovation Ontario Research Fund (MRM). Salary support was provided by a CIHR New Investigator Award (MRM).

Physiotherapy Canada **2020; 72(2); 112–121 doi:10.3138/ptc-2018-0110**

more flexible and clinically useful theory and analytic approach known as generalizability theory (GT), and highlight the latter's advantage using a clinical example.

RELIABILITY

CTT states that a measured value is equal to the (unknown) true value plus a single source of random error. Applied in a reliability context, the true score is the conceptual value that would be obtained if a theoretically infinite number of measurements were performed and averaged within a truly unchanged object of measurement (e.g., a patient in the clinical setting): measured score = true score + error.

The error score is assumed to be random, uncorrelated with the true score and uncorrelated with the errors of other objects of measurement. When sampled over an infinite population of examinees, the average error would be zero. There are two types of reliability, relative and absolute. Relative reliability is quantified in terms of a ratio of variances and often expressed as an intra-class correlation coefficient (ICC). Specifically, reliability is quantified as the true score variance divided by the total variance:

$$\text{Reliability coefficient (ICC)} = \frac{\sigma_{\mathrm{T}}^2}{\sigma_{\mathrm{T}}^2 + \sigma_{\mathrm{E}}^2}, \tag{1}$$

where σ_{T}^2 is the true score variance and σ_{E}^2 is the error variance.

Absolute reliability is often quantified using the standard error of measurement (*SEM*), which is determined by taking the square root of the error variance:

$$SEM = \sqrt{\sigma_{\mathrm{E}}^2}. \tag{2}$$

Although this conceptualization is simple, a consequence of its single source of measurement error necessitates multiple study designs and analyses resulting in multiple estimates of reliability for a given measure. For example, an investigator would have to design different studies to estimate the inter-trial, inter-day, and inter-rater reliabilities of a measure in a specific context. Conceivably, each of these studies could produce a different estimate of reliability owing to a different estimate of measurement error. An important consequence of these distinct estimates of error variance is that they limit one's ability to directly implement the optimal measurement strategy to obtain a clinically acceptable amount of measurement error. By *clinically acceptable,* we mean that the measurement error is sufficiently small as to not interfere with the interpretation of the measured value and any subsequent decisions and actions based on that value.

We define a measurement strategy as the number and type of measurements that a clinician averages to obtain the reported value. For example, a clinician treating patients with Parkinson's disease may be interested in obtaining a sufficiently reliable estimate of a patient's timed up-and-go (TUG) test time. The clinician must debate whether it is better to average trials obtained on the same day or those obtained on different days. The results from separate inter-trial and inter-day reliability studies cannot answer this question satisfactorily.

GENERALIZABILITY

Unlike CTT, GT allows for the simultaneous assessment of multiple sources of measurement error. GT enables researchers to estimate many sources of variance concurrently and thus directly identify the source or sources that contribute most to measurement error. With this information, clinicians and researchers can determine the best measurement strategy for minimizing measurement error.

Applying GT requires investigators to define a set of measurement conditions that define the universe of admissible observations. This universe should identify all sources of measurement error that an investigator deems relevant. Within the universe, each source of error is labelled a *facet,* and the participants (clinical population) investigated are labelled the *objects of measurement.* Returning to the Parkinson's disease example, the facets would be trials and days, and the objects of measurement would be persons with Parkinson's disease. The universe score can be conceptualized as the mean score for an object of measurement over all conditions defined in the universe of generalization. The universe score is similar to CTT's true score; however, unlike CTT, in which an object of measurement has a single true score regardless of the type of reliability being assessed, according to GT, an object of measurement's universe score is facet dependent. Including different facets will produce different universe scores for an object of measurement.

GENERALIZABILITY COEFFICIENTS

Two types of relative reliability coefficients can be calculated using GT: the dependability coefficient and the generalizability coefficient. Dependability is used to determine the reliability of absolute measurements (actual measured values), and generalizability is used to determine the reliability of the rank order of measured values (rank of a person's score among others). These definitions and terms are consistent with the two forms of ICC called *absolute agreement* and *consistency of agreement.* In most clinical and research contexts relevant to physiotherapy, such as evaluating outcomes among patients with Parkinson's disease, researchers are interested in making comparisons with some threshold value and are therefore interested in absolute measurements and the coinciding dependability coefficient.

Both dependability and generalizability are forms of ICC that are calculated as a ratio of variances. We obtain the general form of generalizability coefficients by replacing true score variance from Equation 1 with the universe score variance (σ^2_{U}).

$$\text{Generalizability coefficients: ICC} = \frac{\sigma^2_{\mathrm{U}}}{\sigma^2_{\mathrm{U}} + \sigma^2_{\mathrm{E}}}. \tag{3}$$

The ICC ratio identifies the proportion of total variance accounted for by the objects of measurement (e.g., subjects or participants). If the proportion of participant variance is high, this indicates that the measure is adept at discriminating among participants within the defined universe (i.e., context of interest). The difference between dependability and generalizability comes from the rules used to calculate the universe score and error variances.[1–3]

The *SEM* that coincides with a dependability, or generalizability, coefficient is the square root of the associated error variance (i.e., the same as CTT). In this article, we focus on dependability. In calculating dependability, we determine its associated error variance, called the *absolute error variance,* and calculate the coinciding SEM.

METHODS

There are two steps in any GT analysis. First, we perform a G-study (generalizability study), which is followed by a D-study (decision study). These names should not be confused with the GT coefficients (generalizability and dependability) or with the overall framework of GT itself. The G-study and D-study are outlined next.

G-study

The purpose of a G-study is to determine the variances of as many sources of measurement error (facets) as are relevant and feasible. To obtain these variances, a study must be designed and data collected. For the trial and day facets in the Parkinson's disease example, the most straightforward design is a fully crossed design, in which one collects a specified number of trials on every participant over a specified number of days. The number of trials and days collected must each be at least two. When deciding on the number of collections, it is important that the object of measurement remain unchanged: learning and fatigue should be avoided. Learning may be avoided by providing familiarization sessions before collecting data. Fatigue will likely occur when an excessive number of trials are performed; therefore, researchers should pilot protocols to avoid the progressive decline in muscle performance attributable to fatigue.

Once data have been collected, variance components are calculated. A variance component is the variance for a particular factor in a model; the sum of the variance components is equivalent to a model's total variance. Variance components are acquired by performing an analysis of variance (ANOVA). In the Parkinson's disease example, the dependent variable is the outcome of interest, and predictors are (1) the object of measurement, (2) the facets, and (3) all relevant interactions. The resulting statistical model is as follows:

$$y_{ijk} = u + p_i + t_j + d_k + pt_{ij} + pd_{ik} + td_{jk} + e_{ijk}, \tag{4}$$

in which y_{ijk} is the dependent variable (time to complete the TUG test), u is the mean of all TUG measurements, p_i is the participant factor, t_j is the trial factor, d_k is the day factor, pt_{ij} is the Participant × Trial interaction, pd_{ik} is the Participant × Day interaction, td_{jk} is the Trial × Day interaction, and e_{ijk} is the residual (error).

Using the mean squares calculated in the table (see the Appendix) and the number of levels for each factor, we can calculate the variance components for each term in the model. Variance abbreviations are as follows: σ^2_p = participant variance, σ^2_t = trial variance, σ^2_d = day variance, σ^2_{pt} = Participant × Trial interaction variance, σ^2_{pd} = Participant × Day interaction variance, σ^2_{td} = Trial × Day interaction variance, and σ^2_e = error variance.

To allow easy interpretation, these values are often presented as a percentage of the total variance (sum of variance components). At this stage, it is now possible to examine the variance components and identify the facets that have the greatest sources of error. With this information, we move on to the D-study.

D-study

D-studies focus on identifying the optimal measurement strategy for decision making within the universe that the researcher or clinician wishes to generalize. First, we define the conditions (universe) over which we want to generalize measurements. For the Parkinson's disease example, these are trials and days. That is, we want to be able to determine how well a measurement represents other measurements taken on different days or in different trials. However, it is possible to generalize over a subset of the universe defined in the G-study; we touch on this later.

Knowing the conditions over which we want to generalize, we determine the coefficient or coefficients of interest (dependability or generalizability). We focus on absolute agreement and the dependability coefficient. To calculate dependability, the appropriate universe score variance (σ^2_{U}) and error variance (σ^2_{E}) are calculated.

For this defined universe, the universe score variance is equivalent to the object of measurement (e.g., participant) variance:

$$\sigma^2_{\mathrm{U}} = \sigma^2_p. \tag{5}$$

The dependability-specific error variance, the absolute error variance σ^2_{absE}, for this defined universe is computed as

$$\sigma^2_{\mathrm{absE}} = \frac{\sigma^2_t}{n_t} + \frac{\sigma^2_d}{n_d} + \frac{\sigma^2_{pt}}{n_t} + \frac{\sigma^2_{pd}}{n_d} + \frac{\sigma^2_{td}}{n_t n_d} + \frac{\sigma^2_e}{n_t n_d}. \tag{6}$$

In Equation 6, n_t and n_d are the number of trials (n_t) and days (n_d) over which the outcome measurement (TUG test) is averaged. The calculated σ^2_{absE} is specific to the n_t and n_d input and identifies the σ^2_{absE} for that measurement strategy. The calculated σ^2_{absE} is the sum of the measurement strategy-specific variances for all facets. These are the measurement strategy specific variances because each variance component is divided by the number of trials or days used in the specified strategy. The variance components are divided by the number of measurements (n) because variance is reduced by a factor of n when averaging is used to improve measurement consistency.[4,5] The rule of dividing variance components by the number of averaged measurements for that component is based on the central limit theorem.[5] For this reason, we can change n to any theoretical number (even a number greater than the number we collected) knowing that the variance components will decrease predictably. An inherent benefit of the D-study is that we can calculate the σ^2_{absE} and therefore dependability for any specified measurement strategy (averaging over any specified n_d and n_t).

With the calculated σ^2_{U} and σ^2_{absE}, dependability is computed as

$$dependability = \frac{\sigma^2_{\mathrm{U}}}{\sigma^2_{\mathrm{U}} + \sigma^2_{\mathrm{absE}}}. \tag{7}$$

Equation 7 is consistent with the generic ICC calculation. The coinciding *SEM* is calculated by taking the square root of the σ^2_{absE}.

$$SEM = \sqrt{\sigma^2_{\mathrm{absE}}}. \tag{8}$$

Using individual variances computed from the G-study as well as dependability and the SEM for different measurement strategies (e.g., averaging over different combinations of trials and days), we begin to determine the optimal strategy. First, we critically assess the variance components from the G-study. The facets with the highest variance contribute the most error. Averaging more measurements over these facets will reduce their variability, thus reducing the overall error variance. A lower error variance results in a lower SEM and higher dependability (see Equations 7 and 8). Reducing variability in facets with the highest variances will have the greatest impact. We can more explicitly test and view this result by choosing a range of trials and days (e.g., one to five trials and 1–5 days) for which we are interested in determining the measurement properties. We can then calculate dependability and the *SEM* for all combinations and interpret them graphically or in a table. Graphical and table results are demonstrated next.

CLINICAL EXAMPLE – KNEE EXTENSION POWER

In this section, we provide an example of a GT analysis on knee extension power using synthetic data. The dataset is available as an online-only supplement that includes three trials collected on each of 2 days for 10 participants.

G-study

We began by determining the relevant sources of measurement error and designing a study. The facets of interest were days and trials, and our design mandated three trials on each of 2 days.

The acquired data were fit to the ANOVA model (Equation 4). The dependent variable in the model was knee extension power, and the factors were the same as in Equation 4. The ANOVA model fit using Stata, Version 13.1 (StataCorp LP, College Station, TX) is provided in Table 1. Using the ANOVA mean squares and levels of each factor, the variance components for each factor were calculated using equations from the Appendix and are presented in Table 2.

The interpretation of the variance components is as follows. Each of the variances σ^2_t, σ^2_d, and σ^2_{dt} accounts for 0% of the variance, indicating that there is no systematic difference among the trials or among the days. Zero variance for the interaction (σ^2_{dt}) indicates that the ordered trial means between days are similar. In contrast, σ^2_p accounts for almost 94% of the variance. The high relative variance attributed to the object of measurement indicates that there is relatively small variability remaining for the sources of error. In Equation 7, the dependability coefficient is calculated as the σ^2_{U} divided by the total variance ($\sigma^2_{\mathrm{U}} + \sigma^2_{\mathrm{absE}}$). In Equation 8, the universe score variance is equal to participant variance. From these, our dependability is as follows:

$$dependability = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{absE}}. \tag{9}$$

Without calculating the absolute error (σ^2_{absE}), we know that all other variances are small compared with σ^2_p, indicating that we will have high dependability.

When viewing the other components, σ^2_{pd} accounts for approximately 4% of the variance, which is two to three times the variance for σ^2_{pt} (1.3%) and σ^2_e (1.5%). The σ^2_{pd} variance is indicative of variability from one day to the other, for a particular participant (i.e., some participants performed better on the first versus second day; some better on the second versus first day; others had no

Table 1 Analysis of Variance Results from Fitting Collected Data to the Described Model

Factor	Mean square	*F*	*df*	*p*-value > *F*
Whole model	30,745.76	86.86	41	< 0.001
Participant	135,001.46	381.39	9	< 0.001
Trial	274.40	0.78	2	0.478
Day	260.42	0.74	1	0.400
Participant × Trial	955.68	2.70	18	0.020
Participant × Day	3,053.05	8.63	9	< 0.001
Day × Trial	37.07	0.10	2	0.901
Error	353.97	–	18	–

Notes: The model included 60 observations from 10 participants. Dash indicates no data.

Table 2 Non-Negative Variance for Each Predictor of Peak Knee Extension Power and Its Normalized Variance, Presented as a Percentage of the Total Variance

Symbol	Facet	Variance	Normalized variance (%)
σ_p^2	Participant	21,891.1	93.37
σ_t^2	Trial	0.0	0.00
σ_d^2	Day	0.0	0.00
σ_{pt}^2	Participant × Trial	300.9	1.28
σ_{pd}^2	Participant × Day	899.7	3.84
σ_{dt}^2	Day × Trial	0.0	0.00
σ_e^2	Error	354.0	1.51

difference between the days). Similarly, the interpretation of the σ_{pt}^2 is that individual participants had different trends between trials. In the absolute error calculation (Equation 6), the practice of averaging data can be used to reduce these variances. From Equation 6, averaging over days will reduce variability for the two biggest sources of variance (σ_{pd}^2, σ_e^2) and will thus likely yield the greatest improvement in reliability.

D-study

Next, we determine over which universe to generalize and how to calculate coefficients. The primary purpose is to find out how to obtain the most reliable knee extension power measurements from all relevant protocols. For this reason, we generalize over trials and days, as described for the Parkinson's disease example. We are interested in the reliability of absolute measurements; therefore, we use dependability and its associated *SEM*. Finally, we cover calculating coefficients when we are interested only in generalizing over either trials or days (not both) – these coefficients are analogous to CTTs' inter-trial and inter-day reliability.

To calculate the dependability coefficient and the *SEM* for a single trial ($n_t = 1$) on a single day ($n_d = 1$), we first determine that, according to Equation 5 and variances from Table 2, the universe score variance is $\sigma_{\mathrm{U}}^2 = \sigma_p^2 = 21,891.1$. Next, we input $n_t = 1$, $n_d = 1$, and the appropriate variances (from Table 2) into Equation 6 to calculate the absolute error variance:

$$\sigma_{\mathrm{absE}}^2 = \frac{0.0}{1} + \frac{0.0}{1} + \frac{300.9}{1} + \frac{899.7}{1} + \frac{0.0}{1 \times 1} + \frac{354.0}{1 \times 1}.$$

$$\sigma_{\mathrm{absE}}^2 = 1,555.6.$$

Finally, dependability and the *SEM* are calculated by inputting σ_{U}^2 and σ_{absE}^2 into Equations 7 and 8:

$$dependability = \frac{21,891.1}{21,891.1 + 1,555.6}.$$

$$dependability = 0.934.$$

$$SEM = \sqrt{1,557.3}.$$

$$SEM = 39.44N - m/s.$$

This dependability coefficient indicates how well we can generalize a single trial from one day to any trial from any day.

$$inter - trial\ dependability = \frac{\sigma_p^2 + \frac{\sigma_{pd}^2}{n_d}}{\sigma_p^2 + \frac{\sigma_{pd}^2}{n_d} + \frac{\sigma_t^2}{n_t} + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{td}^2}{n_t n_d} + \frac{\sigma_e^2}{n_t n_d}}. \quad (10)$$

The calculated dependability for a single trial on a single day provides valuable information about the reproducibility of measurements. Beyond that, we can determine the actual reliability of various measurement strategies by changing n_t and n_d in the error variance calculation (Equation 6). Table 3 includes calculated dependability and SEM values for measurement strategies that include averaging over all combinations of one to five trials and 1–5 days. Both n_t and n_d can be greater than the number of collected trials (3) and days (2) from the G-study because we are using the central limit theorem to identify what the variance components will be when we use averaging to improve measurement consistency.

The resulting dependability and SEM values for all calculated measurement strategies are presented visually in Figures 1 and 2. The stacked nature of the individual lines clearly demonstrates that higher reliability can be obtained by averaging over days versus averaging over trials. This highlights the fact that when attempting to identify change, variability among days should be accounted for if a true change is to be detected.

DISCUSSION

Classical test theory equivalents

Thus far, the presented dependability coefficients have commented on the reliability over all measurement conditions within the defined universe. However, it is useful to identify a specific measure of inter-day and inter-trial reliability. The GT equivalent of traditional inter-trial or inter-day reliability can be calculated using additional rules for calculating universe score variance and absolute error variance.[2] Following these rules, the inter-trial reliability equivalent is calculated using Equation 10, and the inter-day equivalent using Equation 11.

$$inter - day\ dependability = \frac{\sigma_p^2 + \frac{\sigma_{pt}^2}{n_t}}{\sigma_p^2 + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_d^2}{n_t} + \frac{\sigma_{pd}^2}{n_d} + \frac{\sigma_{td}^2}{n_t n_d} + \frac{\sigma_e^2}{n_t n_d}}. \quad (11)$$

These coefficients refer to the ability to generalize across a single facet. The inter-trial dependability coefficient is interpreted as “the extent to which I can generalize from one trial to another,”[2] and the inter-day coefficient is “the extent to which I can generalize from one day to another.” For the current example, if we assume $n_t = 1$ and $n_d = 1$, we find

$$inter - trial\ dependability = \frac{21,891.1 + \frac{899.7}{1}}{21,891.1 + \frac{899.7}{1} + \frac{0}{1} + \frac{300.9}{1} + \frac{0}{1\times1} + \frac{354.0}{1\times1}}.$$

$$inter - trial\ dependability = 0.972.$$

$$inter - day\ dependability = \frac{21,891.1 + \frac{300.9}{1}}{21,891.1 + \frac{300.9}{1} + \frac{0}{1} + \frac{899.7}{1} + \frac{0}{1\times1} + \frac{354.0}{1\times1}}.$$

$$inter - day\ dependability = 0.947.$$

This exercise further highlights the fact that there is greater measurement error across days than across trials.

In the following, we compare overall coefficients as well as trial- and day-specific coefficients for dependability and the SEM with their traditional ICC and SEM counterparts. Table 4 includes Shrout and Fleiss Type 1,1 and Type 2,1

Table 3 Dependability Coefficient and SEM for Measurement Strategies That Include Averaging Overall Combinations of One to Five Trials and 1–5 Days

	Trials				
Days	1	2	3	4	5
Dependability					
1	0.934	0.947	0.951	0.954	0.955
2	0.960	0.970	0.973	0.975	0.976
3	0.968	0.977	0.980	0.982	0.983
4	0.973	0.981	0.984	0.985	0.986
5	0.975	0.984	0.986	0.988	0.989
SEM, N-m/s					
1	39.43	35.03	33.44	32.61	32.10
2	30.46	26.24	24.68	23.86	23.35
3	26.81	22.57	20.96	20.11	19.59
4	24.78	20.48	18.83	17.95	17.40
5	23.49	19.12	17.43	16.52	15.95

SEM = standard error of measurement.

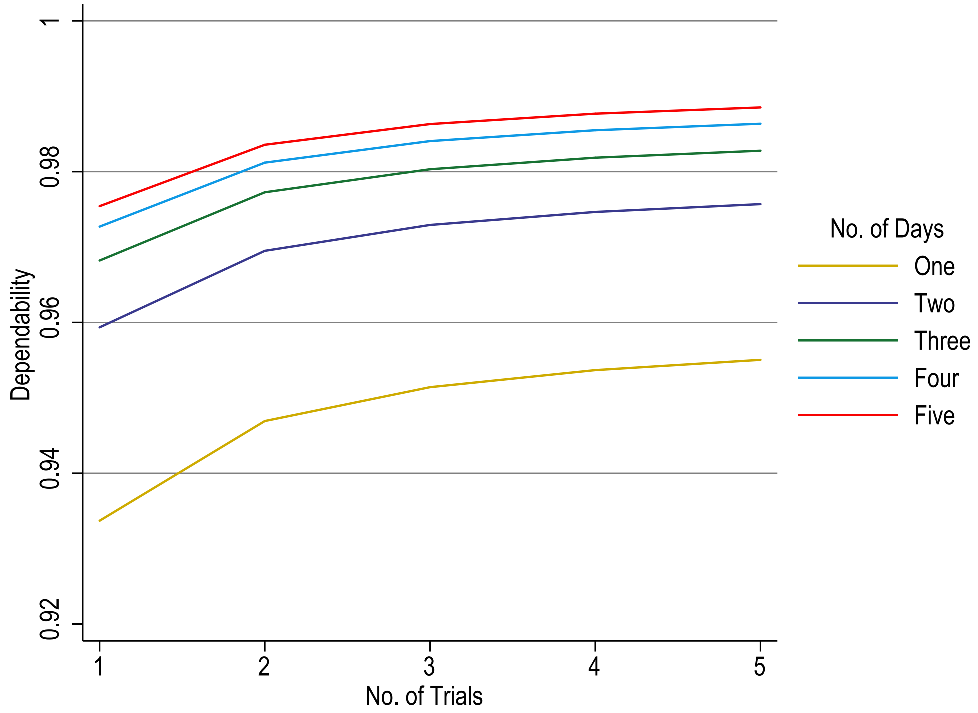

Figure 1 Dependability of knee extension power measurements presented for measurement strategies that include averaging over one to five trials and 1–5 days.
Notes: The *x*-axis identifies the number of trials averaged. Individual lines represent averaging over different numbers of days.

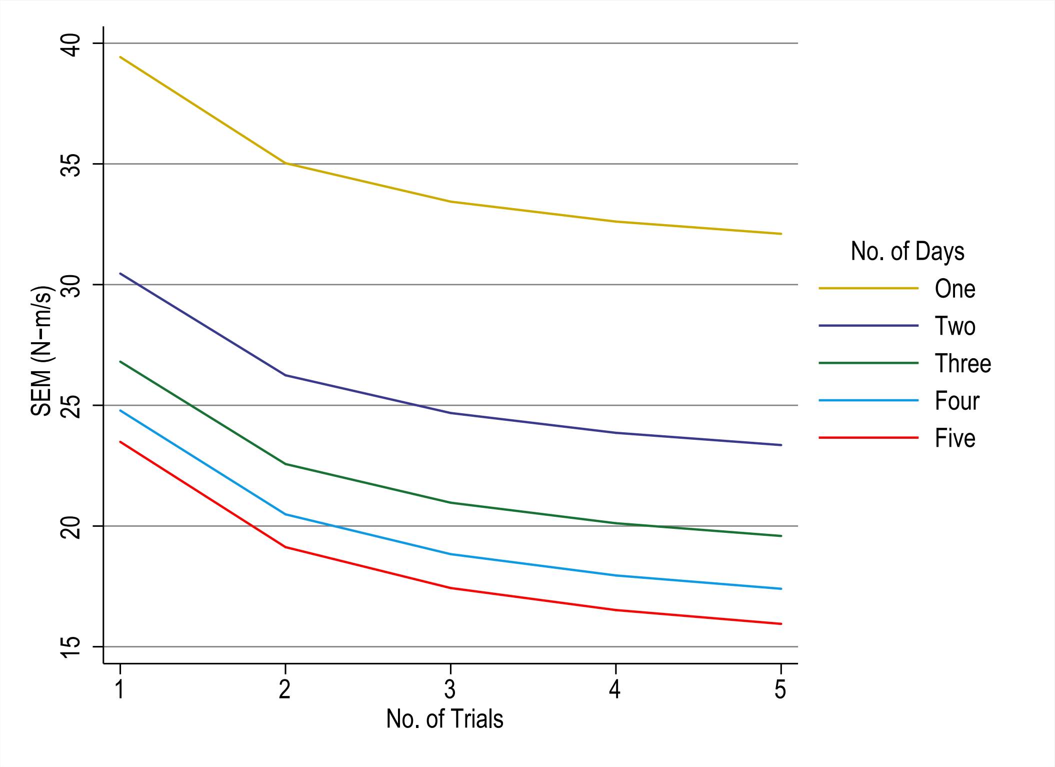

Figure 2 SEM of knee extension power measurements presented for measurement strategies that include averaging over one to five trials and 1–5 days.
Notes: The *x*-axis identifies the number of trials averaged. Individual lines represent averaging over different numbers of days.
SEM = standard error of measurement; N-m/s = newton metres per second.

ICCs and the corresponding SEM for inter-trial and inter-day reliability. Type 1,1 ICCs follow CTT, and Type 2,1 ICCs are equivalent to a dependability coefficient for a single facet and a single measurement ($n = 1$).

Table 4 also includes a summary of the coefficients presented in the GT example. Using GT, we obtained three sets of coefficients: overall reliability, reliability across trials, and reliability across days. From the same data, using ICCs, we obtained two separate estimates of inter-trial reliability, three estimates for inter-day reliability, and no representation of overall reliability. Multiple estimates for inter-trial and inter-day reliability were obtained because CTT studies cannot partition the variances between days and trials simultaneously. On closer inspection, the GT equivalents to inter-trial and inter-day reliability are approximately the average of the ICC estimates obtained using Type 1,1 and Type 2,1 ICCs. GT accounts for all included data from all days and trials, and it appropriately partitions the variances, providing a more robust estimate of reliability.

Using either Type 1,1 or Type 2,1 ICCs, it is likely that a study would obtain and report from only a single assessment and thus likely under- or overestimate reliability. Moreover, when assessing SEMs, it can be seen that CTT underestimates measurement error, leaving clinicians more confident in their measurement than they should be. GT also comes with the added insight into how a single measurement generalizes over trials and days, a metric that is useful in real-world applications.

Standard error of measurement insights

Missing from this analysis is how to apply these findings to clinical practice. Two important clinical questions when assessing a patient are (1) how confident can I be in a measured value and (2) on reassessment, has this patient changed? In the context of a generalizability approach, the term *measured value* is replaced with the average value obtained from the specific measurement strategy. To provide clinicians with an estimate of confidence in their measurements, the SEMs can be used to calculate a CI for any measurement strategy. CIs are obtained by multiplying the SEM by the z-score of desired confidence. From the presented example, a 90% CI for the average of four trials from 1 day is calculated as

$$CI = measurement \pm 1.645 \times SEM(4\ trials, 1\ occasion)$$

$$CI = measurement \pm 1.645 \times 32.61N - m/s$$

$$CI = measurement \pm 53.64N - m/s$$

where 1.645 is the z-score associated with a 90% confidence level. For example, suppose this measurement strategy yielded a power of 300 N-m/s for a patient. The clinician could be 90% confident that this patient's true power was 300 ± 53.64 N-m/s.

To answer the second question, a multiple of the SEM could be used to calculate the minimum detectable change (MDC) at a given confidence level. The MDC is calculated by multiplying the SEM, the square root of 2 (this acknowledges that there is error associated with the two measured values being compared), and the z-score for the desired confidence level. Therefore, again, for the average of four trials from 1 day, the MDC at a 90% confidence level can be calculated as

$$MDC_{90} = SEM(4\ trials, 1\ occasion) \times \sqrt{2} \times 1.645$$

$$MDC_{90} = 32.61N - m/s \times \sqrt{2} \times 1.645$$

$$MDC_{90} = 75.86N - m/s$$

The interpretation of MDC$_{90}$ is that 90% of truly unchanged patients would display random fluctuations within the bounds defined by MDC$_{90}$.

The CI and MDC are valuable tools clinicians can use to assess changes in outcome measures. To reiterate, the CI can be used to identify the confidence in the average value obtained from the measurement strategy. However, the MDC can be applied when assessing a change between two separate measurements of the same outcome. The MDC should be used in such instances because it accounts for error associated with both measurements.

Calculating the CI and MDC$_{90}$ highlights something that is not identifiable from the coinciding ICC (0.934) and not obvious from the SEM (32.61 N-m/s): the error in measuring knee extension power is large. As shown in the MDC example calculation, if a therapist were to average over four collected trials on a single day, the SEM of 32.61 N-m/s would yield an MDC$_{90}$ of 75.86 N-m/s. This means that the patient would have to change by more than 75.86 N-m/s between days for the therapist to be reasonably certain that the change was real. When compared with the average knee extension power of the knee osteoarthritis sample dataset included in the online-only supplement (297.7 watts), that equates to a 25.4% change.

In comparison, if we use the SEM of the average measurement from two trials taken over 2 days (26.24 N-m/s; see Table 3), the calculated MDC$_{90}$ is 61.04 N-m/s (20.5 % of the mean). Averaging the same number of total measurements (four) over 2 days reduces the SEM and, as a result, reduces the MDC$_{90}$, improving the ability to detect change. In this example, averaging over 2 days improves the ability to detect change because more variance is attributed to facets that include days (mainly σ^2_{pd}) than facets that include trials (mainly σ^2_{pt}). This highlights the fact that averaging over facets with greater variances has the greatest overall impact on improving reliability. In general, these findings highlight that even after averaging over 2 days, measurement error and the MDC$_{90}$ is large

Table 4 Summary of Variances, SEM, and ICCs for Inter-Trial and Inter-Day Reliability, Calculated Using CTT and for Inter-Trial, Inter-Day, and Overall Reliability Using GT

			Classical reliability		Typically reported reliability (single-facet generalizability)		
Measure being summarized	Variance source	Model number	Inter-trial	Inter-day	Inter-trial	Inter-day	Generalizability
Variances							
	Participants						
		1	22,539.16	23,437.64	22,525.73	23,392.00	21,891.12
		2	23,075.73	21,456.03	23,052.89	21,393.58	–
		3	–	21,853.56	–	21,787.33	–
	Trials						
		1	–	–	0	–	0
		2	–	–	0	–	–
	Days						
		1	–	–	–	0	0
		2	–	–	–	0	–
		3	–	–	–	0	–
	PT		–	–	–	–	300.85
	PV		–	–	–	–	899.69
	TV		–	–	–	–	0
	Error						
		1	484.90	951.65	516.19	1,042.94	353.97
		2	724.93	1,268.65	793.47	1,387.56	–
		3	–	1,198.05	–	1,330.49	–
SEM							
		1	22.02	30.85	22.72	32.29	Overall 39.43
		2	26.92	35.61	28.17	37.25	Inter-trial 25.59
		3	–	34.61	–	36.48	Inter-day 35.41
$ICC_{1,1}$							
		1	0.98	0.96	–	–	–
		2	0.97	0.94	–	–	–
		3	–	0.95	–	–	–
$ICC_{2,1}$							
		1	–	–	0.98	0.96	–
		2	–	–	0.97	0.94	–
		3	–	–	–	0.95	–
ICC							
			–	–	–	–	Overall 0.93
			–	–	–	–	Inter-trial 0.97
			–	–	–	–	Inter-day 0.95

Note: Dash indicates no data.

SEM = standard error of measurement; ICCs = intra-class correlation coefficients; CTT = classical test theory; GT = generalizability theory; PT = Participant × Trial interaction; PV = Participant × Visit interaction; TV = Trial × Visit interaction.

(> 20%), making it difficult to identify individual changes in knee extension power. This information is critical to allow clinicians to make appropriate decisions based on their measurements.

CONCLUSION

When devising a measurement strategy for clinical practice or research, it is important that all potential sources of measurement error be identified so that an optimal strategy can be implemented. GT enables researchers to systematically analyze and compare the relevant sources of measurement error through their variances. Beyond allowing for general comparisons, GT enables researchers to determine the relative reliability in the form of a dependability coefficient, as well as the absolute reliability in the form of the SEM, for any measurement strategy that they choose to use. Using GT enables researchers to develop a well-rounded interpretation of measurement error for clinical outcomes, thereby enabling them to identify optimal measurement strategies and provide information about measurement errors that can be conveyed to clinicians to aid in shaping their clinical decisions. For those seeking to use GT, a number of other resources may be of interest.[1–4]

KEY MESSAGES

What is already known on this topic

The frameworks of reliability and generalizability theory (GT) have been reported for decades. The benefits of GT have been thoroughly outlined and described, particularly by Robert L. Brennan.[3] Its major benefits have been to identify various sources of measurement error and to use this information to maximize the reliability or generalizability of measurement.

What this study adds

This study adds an explanation of GT and an example of its application that is useful to clinical researchers. We explain the background and terminology of GT in a way that is relevant to physiotherapy and include an explicit example of the theory using relevant data (knee extension power). The dataset used for this example is provided to allow readers to follow along and to serve as a resource for teaching measurement and generalizability theories to rehabilitation scientists.

REFERENCES

1. Briesch AM, Swaminathan H, Welsh M, et al. Generalizability theory: a practical guide to study design, implementation, and interpretation. J Sch Psychol. 2014;52(1):13–35. https://doi.org/10.1016/j.jsp.2013.11.008. Medline:24495492
2. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach. 2012;34(11):960–92. https://doi.org/10.3109/0142159X.2012.703791. Medline:23140303
3. Brennan RL. Generalizability theory. New York: Springer; 2001. https://doi.org/10.1007/978-1-4757-3456-0.
4. Shavelson RJ, Webb NM. Generalizability theory. In: Green JL, Camilli G, Elmore PB, editors. Handbook of complementary methods in education research. Washington (DC): American Educational Research Association; 2006. p. 309–22.
5. Li M, Shavelson RJ, Yin Y, et al. Generalizability theory. In: Cautin RL, Lilienfeld SO, editors. Encyclopedia of clinical psychology. Hoboken (NJ): Wiley; 2015. p. 1–19. https://doi.org/10.1002/9781118625392.wbecp352.
6. Searle SR, Casella G, McCulloch CE. Variance components. Hoboken (NJ): Wiley; 2006.
7. Kleinbaum DG, Kupper LL, Nizam A, et al. Applied regression analysis and other multivariable methods. 5th ed. Boston: Cengage Learning; 2013.

APPENDIX

In this appendix, we provide the formulas for calculating the variance components from an analysis of variance model (ANOVA). We provide formulas for a three-way fully crossed ANOVA model – that is, a model that includes three factors – participant (*p*), trial (*t*), and day (*d*), all two-way interactions, and a residual error (*e*) term. The factors do not necessarily have to be *p, t,* and *d;* these equations apply to any fully crossed three-way ANOVA. The calculations are derived from the mean squares and the levels of each factor.

We should note that when calculating variances from an ANOVA table, it is possible to encounter negative variances. However, variances are not negative.[6] In the event of a negative variance component, a common way to proceed is to assign it a variance of 0. We have used this methodology in the current example. A more rigorous approach would be to use other methods of fitting the model and obtaining variance components, such as using a mixed effects model that uses maximum likelihood (ML) or restricted maximum likelihood (REML) to estimate the variances. Explanations of random effects models, ML, and REML can be found in common statistics textbooks.[7]

Equations for calculating variance components from a three-way fully crossed ANOVA model

ANOVA factor	Variance component calculation
p	$\frac{MS_p - MS_{pt} - MS_{pd} + MS_e}{n_t n_d}$
t	$\frac{MS_t - MS_{pt} - MS_{td} + MS_e}{n_p n_d}$
d	$\frac{MS_d - MS_{pd} - MS_{td} + MS_e}{n_p n_t}$
pt	$\frac{MS_{pt} - MS_e}{n_d}$
pd	$\frac{MS_{pd} - MS_e}{n_t}$
td	$\frac{MS_{td} - MS_e}{n_p}$
e	MS_e

ANOVA = analysis of variance; MS_x = mean square for factor *x*; n_x = number of levels for factor *x*.