Behav Res (2018) 50:2162–2172
https://doi.org/10.3758/s13428-017-0983-6

BRIEF COMMUNICATION

Trajectory analysis of discrete goal-directed pointing movements: How many trials are needed for reliable data?

Jarrod Blinch[1] · Youngdeok Kim[1] · Romeo Chua[2]

Published online: 7 December 2017

Abstract A powerful tool in motor behavior research is trajectory analysis of discrete goal-directed pointing movements. The purpose of the present analysis was to estimate the minimum number of trials per participant required to achieve the conventional level of reliability for trajectory analysis. We analyzed basic measurements of movement and three common methods of trajectory analysis within the framework of generalizability theory. Generalizability studies were used to decompose the total variance of these variables into the percent contributions from person, trial, and the person-by-trial interaction. Decision studies were then used to determine the minimum number of trials required to achieve the conventional level of reliability. The number of trials per participant needed for reliable data of discrete goal-directed pointing movements depended on the dependent variable—for example, reaction times required six or ten trials, movement times required three trials, and constant error required 47 trials. For trajectory analysis, ten or fewer trials were required for reliable dependent variables during the first half of the movement (up to peak velocity or 70% of the displacement). The number of trials required for the second half of the movement rapidly increased to 47 trials at movement termination. This increase in the number of trials required for reliable analysis of the second half of the movement was indicative of online control. Finally, correlation analysis was performed with simulated correlations on subsets of trials, and all 32 trials were required. However, 18 trials might be used without a practically significant change in the correlations.

Keywords Generalizability theory · Kinematic landmarks · Variability analysis · Time-normalized profiles · Correlation analysis

A powerful tool in motor behavior research is trajectory analysis.[1] Trajectory analysis involves evaluating movement progression in space and time. This spatiotemporal analysis dates back to at least the advent of chronophotography in the Victorian era by people like Eadweard Muybridge and Étienne-Jules Marey. Other important landmarks in the technology behind trajectory analysis include mechanical methods (e.g., Woodworth, 1899), film analysis (e.g., Bernstein, 1967; Jeannerod, 1984), and optoelectric motion capture systems (e.g., MacKenzie, Marteniuk, Dugas, Liske, & Eickmeier, 1987; Marteniuk, MacKenzi, Jeannerod, Athenes, & Dugas, 1987). Trajectory analysis has been applied to all sorts of human movements including gait, balance, and reach-to-grasp and pointing movements. Identifying the spatiotemporal properties of a movement allows us to infer how the nervous and musculoskeletal systems produced the movement.

We focused on the trajectory analysis of discrete goal-directed pointing movements in this article. Much of what is discussed, however, can be applied to the reach component of

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-017-0983-6) contains supplementary material, which is available to authorized users.

✉ Jarrod Blinch
jarrod.blinch@gmail.com

[1] Department of Kinesiology & Sport Management, Texas Tech University, Lubbock, TX, USA

[2] School of Kinesiology, University of British Columbia, Vancouver, British Columbia, Canada

[1] Many review articles have discussed trajectory analysis, or more specifically, the relationship between aspects of the trajectories and cognitive processes. A few examples are Desmurget, Pélisson, Rossetti, and Prablanc (1998); Elliott, Helsen, and Chua (2001); Elliott et al. (2017); Gaveau et al. (2014); and Prablanc, Desmurget, and Gréa (2003).

reach-to-grasp movements and, potentially, to other types of movements. Our results are unlikely to apply to fundamentally different movements, such as reaches with target perturbations, trajectory deviations (e.g., action-dynamics tasks), or motor contagion and imitation. There have been a few articles on the methodology of trajectory analysis for goal-directed movements—for example, filtering to reduce noise in kinematics (Winter, Sidwall, & Hobson, 1974), detecting online control (Khan et al., 2006), quantifying the variability of aiming movements (Hansen, Elliott, & Khan, 2008), robust movement segmentation (Schot, Brenner, & Smeets, 2010), and determining the latency of online corrections (Oostwound Wijdenes, Brenner, & Smeets, 2014). A vital issue that requires further investigation is how many trials are needed for reliable measures of trajectories. This could be estimated by making empirical estimates of the intra-individual variability of trajectory measures across multiple trials. The number of trials per condition in articles with trajectory analysis of goal-directed movements has been inconsistent. We estimated this inconsistency by randomly sampling 13 articles with trajectory analysis from the 127 references in a recent review article on goal-directed reaching movements (Elliott et al., 2017). The minimum number of trials per condition was four (Melmoth, Storoni, Todd, Finlay, & Grant, 2007) and the maximum was 60 (Welsh, Higgins, & Elliott, 2007). The mean number of trials was 20.3, and the 95% confidence interval was [11.9, 28.8]. This inconsistency makes it difficult to estimate how many trials are needed per condition; should one test 12, 20, or 29 trials? Two guidelines that have been passed down to determine how many trials to include are (1) the more trials the better and (2) 12 to 15 good trials per condition might be adequate. Unfortunately, the second guideline was not systematically developed by investigating the reliability of the trajectories based on empirical estimates of intra-individual variability.

According to classical test theory, reliability refers to the degree to which a test score is consistent across repeated observations, under the assumption the true score is the average of observed test scores obtained over an infinite number of repeated observations. The total variance of any observed test score is, therefore, composed of two variance components, the variance of the true score and the variance of the measurement error. The reliability of the test score is quantified as the proportion of true score variance to the total variance (variance of the true score plus measurement error). Thus, reliability decreases when the variance associated with measurement error increases. Generalizability theory (G theory) is an extension of classical test theory through the application of analysis of variance methods to decompose the variance of observed test scores into multiple sources of measurement error, as well as the variance associated with interactions between sources (reviewed by Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Vispoel, Morris, & Kilinc, 2017). Another unique feature of G theory is the estimate of the change in reliability by exploring the various conditions of measurement. This allows one to determine the optimal measurement conditions that maximize reliability.

In the analysis of discrete pointing movements, for example, reaction time is measured for a group of participants on several trials. This is described as a person crossed with trial experiment in the G theory framework. A generalizability study (G study) first estimates the variance contributed by person, trial, and the person-by-trial interaction, which is similar to main effects and interactions in an analysis of variance. The person variance is caused by interindividual differences in reaction time, which is true variance. The trial and person-by-trial interaction variances are the two sources of measurement error. These represent the systematic error associated with intertrial variability and the random error associated with intra-individual variability, respectively. On the basis of the estimated variance components from the G study, a follow-up Decision study (D study) can then estimate the change in reliability by increasing and decreasing the number of trials per participant (described in the Material and method section). G theory has been widely used in the many subdisciplines of kinesiology. Examples include estimating inter- and intrarater variability in movement skill assessment (Kim, Park, & Kang, 2012) and determining the reliability of accelerometry-based activity monitors (Welk, Schaben, & Morrow, 2004), pedometers (Kang, Bjornson, Barreira, Ragan, & Song, 2014), the balance error scoring system (BESS; Broglio, Zhu, Sopiarz, & Park, 2009), center of pressure measures (Doyle, Hsiao-Wecksler, Ragan, & Rosengren, 2007), blood pressure readings (Llabre et al., 1988), isometric force measurement (Roebroeck, Harlaar, & Lankhorst, 1993), physical activity questionnaire for children (PAQ-C; Crocker, Bailey, Faulkner, Kowalski, & McGrath, 1997), and the eating attitudes test (EAT-12; Engelsen & Hagtvet, 1999).

The present analysis applied G theory to the basic measurements of movement (reaction time, movement time, and constant error) and three common methods of trajectory analysis (kinematic landmarks, spatial variability, and time-normalized displacement profiles). We examined the variance components associated with different sources of measurement error (trial and person-by-trial) to determine the optimal measurement conditions to achieve the conventional level of reliability (≥.80; Shrout, 1998) for each dependent variable. The goal was to establish proven guidelines for future research to ensure reliable measurement of discrete goal-directed pointing movements. Knowing how many trials to include will assist in the experimental design and ensuring the measurements are reliable, reproducible, and consistent will improve the experimental quality.

Material and method

The data for the present analysis were previously reported in Blinch et al. (2014). The relevant details from that study are summarized in the following subsections and the present analysis with G theory is detailed.

Participants, apparatus, procedure, and design

The data from the original twenty participants were analyzed. The present analysis focused on the six unimanual movement blocks. These involved making pointing movements to targets (15.3-mm radius) with long (200 mm) or short (100 mm) movement amplitudes. Participants pointed with a stylus and were instructed to "hit the targets as quickly as possible." An infrared emitting diode was placed near the tip of the stylus and its position was recorded with an Optotrak (3020; Northern Digital Inc.) at 250 Hz. Four of the blocks were simple reaction-time conditions. In a simple reaction time task, participants know which movement will be required before the go signal. The movements tested were as follows: *left* arm to a target with a *long* movement amplitude, left short, right long, and right short. The remaining two blocks were two-choice reaction-time conditions. In a choice reaction time task, participants do not know which one of two movements will be required until the go signal. One block tested movements with the left arm (long or short) and the right arm (long or short) was tested in the other block. Every block included 32 test trials of each movement type. All participants were tested on all the blocks; it was a repeated measures design, which G theory refers to as a fully crossed design.

Basic measurements of movement and trajectory analysis

The basic measurements of movement were reaction time, movement time, and constant error, which were calculated for each trial. A microswitch in the stylus was used to determine when the stylus tip was pressed against the table and when it was lifted. This signal was used to calculate reaction time (time from the go signal to stylus lift) and movement time (time from stylus lift to stylus press). The position of the stylus when it was pressed at the end of the movement time was used to calculate constant error; the constant error was the position of the stylus in the primary direction of the movement minus the position of the center of the target.

For trajectory analysis, we focused on the displacement data in the sagittal plane, the primary direction of the movements. These data were filtered with a 20-Hz Butterworth low-pass filter (second order, dual pass). We examined kinematic landmarks throughout movement execution—specifically, the time, magnitude, and position/variability of positive peak acceleration, as well as peak velocity and negative peak acceleration (Fig. 1). The variability at the kinematic landmarks can be used to infer the amount of online control during movements (e.g., Khan & Franks, 2003). The position at the kinematic landmarks is used for the same purpose (e.g., Carlton et al., 1984; Gordon & Ghez, 1987; Heath, Westwood, & Binsted, 2004), but the positions at each kinematic landmark are first correlated with the positions at movement termination. (Both techniques are reviewed by Khan et al., 2006.) For the analysis of the time-normalized displacement profiles, the displacement during the movement time of each trial was interpolated into 100 frames. This converted the movement from time in milliseconds to percent time. This allowed trials with shorter or longer movement times to be averaged together from movement initiation to termination. The positions from 10% to 100% time, in increments of 10%, were then extracted.

Data analysis

To follow is a brief explanation of the G theory design used in the present analysis and the equations used in the G and D studies. Suppose the dependent variable is reaction time; reaction time can be measured from people (*p*) on many trials (*t*). Any observed single reaction-time score is represented with Eq. 1.1.

$$X_{\mathrm{pt}} = \mu + \nu_{\mathrm{p}} + \nu_{\mathrm{t}} + \nu_{\mathrm{pt}} \tag{1.1}$$

In Eq. 1.1, μ is the grand mean and ν designates the components in the design (person, trial, and the person-by-trial residual effect). If we measure reaction time for all people and all trials, then the total variance of all observed scores is given by Eq. 1.2.

$$\sigma^2\left(X_{\mathrm{pt}}\right) = \sigma^2(p) + \sigma^2(t) + \sigma^2(pt) \tag{1.2}$$

Equation 1.2 shows that the total variance has been decomposed into the random effects variance components of person, trial, and the person-by-trial interaction.

A G study is then run to estimate the variance components: the true variance across people [$\hat{\sigma}^2(p)$; interindividual], the systematic error variance across trials [$\hat{\sigma}^2(t)$; intertrial], and the random error variance of the rank ordering of people across trials [$\hat{\sigma}^2(pt)$; intra-individual]. The estimated variance components are best interpreted as the percentage they contribute to the total estimated variance [$\hat{\sigma}^2(p) + \hat{\sigma}^2(t) + \hat{\sigma}^2(pt)$].

A D study can then be run to estimate the minimum number of trials required per participant to achieve a certain level of reliability. We used the conventional level of reliability of .80 in the present analysis [generalizability coefficient (g) ≥ .80]. The generalizability coefficient[2] is estimated with the following equation:

$$g\ coefficient = \hat{\sigma}^2(p)/\left(\hat{\sigma}^2(p) + \hat{\sigma}^2(pt)/n_{\mathrm{T}}\right), \tag{1.3}$$

[2] D studies can use either the generalizability coefficient or the phi coefficient. The generalizability coefficient measures relative consistency and absolute consistency is measured by the phi coefficient. We used the generalizability coefficient in the present analysis because we were interested in the degree to which the dependent variables maintained their rank across people and trials regardless of the actual scores. Absolute consistency would also consider the degree of consistency of the actual scores. One reason we were not interested in absolute consistency was that we expected that some participants would have shorter reaction time than others.

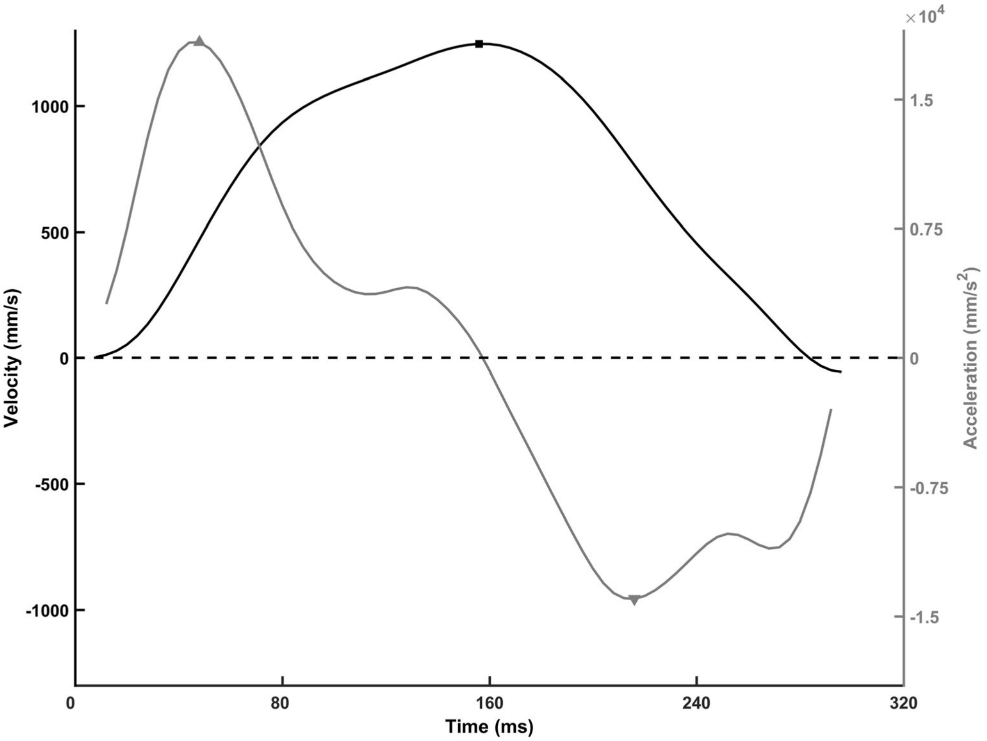

Fig. 1 Velocity (black line) and acceleration (gray line) profiles on a typical trial. These profiles were used to determine the time and magnitude of positive peak acceleration (gray upward triangle), peak velocity (black square), and negative peak acceleration (gray downward triangle)

where n_T is the number of trials per participant. The number of trials begins at one and is increased until the generalizability coefficient is greater than or equal to .80. Note that the number of trials, n_T, can be larger than the actual number of trials in the D study.

G theory can be applied for any dependent variable we can measure on each trial, for example the position at a kinematic landmark. G theory may not be applicable for aggregate dependent variables that are measured over a series of trials, for example, the spatial variability at a kinematic landmark, which is calculated by taking the standard deviation of the positions. This makes the variability a secondary outcome that is dependent on the position data. If G theory analysis shows that the positions at the kinematic landmarks are reliable with a certain number of trials, then the variability at the kinematic landmarks are most likely reliable as well. In the present analysis, we, therefore, applied G theory analysis on the position data and used those results to make conclusions about the reliability of the positions and the variability at the kinematic landmarks.

The data were analyzed with a MATLAB G theory program created by Mushquash and O'Connor (G1.m[3]; 2006). The nfacet1 variable was set to 32 (the number of trials), and the type variable was set to 1 (indicating a single-facet, fully crossed design).

An important issue for G and D studies is the generalizability of the results. The key here is that we chose to have G theory treat the object of measurement (people) and the condition of measurement (trials) as random effects—specifically, in this study, the object of measurement is a subset of any person in the population, and the condition of measurement is a subset of any potential trial (called the infinite universe of trials in generalizability theory). The G and D study results, therefore, generalize to movements with similar measurement conditions, which is detailed in the Discussion section.

The correlation analysis required an entirely different approach than G theory because a correlation cannot be measured for each trial. Our solution was to compare the accuracy of simulated correlations on a subset of trials (three to 30 trials) to the actual correlations with 32 trials. For a subset of 20 trials, for example, we selected a random subset of 20 trials and then calculated the position correlations of positive peak acceleration, peak velocity, and negative peak acceleration with movement termination. This procedure was repeated 100 times and the mean simulated correlations were calculated for each participant. A random subset was selected on each repetition without replacement. We could not perform this analysis for subsets of one, two, or 31 trials (you cannot correlate a single trial, two trials will [almost] always yield a perfect correlation, and there are only 32 random subsets of 31 trials). The simulated correlations for subsets of three to 30 trials were compared to the actual correlations with 32 trials

[3] G1.m is available for download from https://people.ok.ubc.ca/brioconn/gtheory/gtheory.html. Along with the version for MATLAB, there are versions for SPSS, SAS, and R.

with a series of paired-samples *t* tests. The Bonferroni correction was used to control the familywise error rate. These 28 *t* tests were used on the data for positive peak acceleration, peak velocity, and negative peak velocity.

Results

The G and D study results were very similar for the different types of movements in simple and choice reaction time tasks. We, therefore, decided to present the results for the right arm, long distance movements in the choice reaction time task. (The full analysis and all the data are available as supplementary material.) The one exception to this was for the reaction time results, in which the G and D study results were slightly different in choice and simple reaction time tasks. For reaction time, we presented the results for right long movements in choice and simple reaction time tasks. The variability of the D study results across the eight conditions are also presented in the section on the variability of the D study results.

Basic measurements of movement

Reaction time, movement time, and constant error results are shown in Table 1. For reaction time in the choice reaction time task, 39.8% of the total variance was caused by person, the individual differences in reaction time. Only 5.1% of the total variance was caused by trial. This suggested that reaction time, when averaged across participants, remained relatively consistent across the 32 trials. The interaction between person and trial caused 55.1% of the total variance, which indicated the rank ordering of participants differed across trials. A D study estimated that at least six trials per participant were needed to achieve the conventional .80 level of reliability. The results in the simple reaction time task were slightly different. Less variability was caused by person (28.6%), and more variability was caused by the random errors attributed to the person-by-trial interaction (65.4%). This decrease in true variance and increase in random error variance resulted in at least ten trials being needed to achieve the conventional level of reliability.

For movement time, the majority of the total variance (62.5%) was caused by person, very little of the variance (0.2%) was caused by trial, and the remaining variance (37.3%) was caused by the interaction. A D study estimated that only three trials per participant were required to achieve the conventional level of reliability. Interestingly, at least 47 trials were needed to achieve conventional reliability for constant error. This large number of trials was needed because almost all of the total variance (92.1%) was caused by the person-by-trial interaction.

Trajectory analysis

Time and magnitude of the kinematic landmarks The results for the time and magnitude of positive peak acceleration, peak velocity, and negative peak acceleration are shown in Table 2. For the time of the kinematic landmarks, the total variance was split between person and person-by-trial. The time of positive peak acceleration had the largest person-by-trial variance (55.5%), but only five trials were needed to achieve the conventional level of reliability. The time of peak velocity and negative peak acceleration had less person-by-trial variance (35.5% and 42.04%, respectively), and only three trials were needed for conventional reliability. For the magnitude of the kinematic landmarks, about 75% of the total variance was caused by person, and 25% by the person-by-trial interaction. This high percentage of true variance and low percentage of random error variance resulted in only one or two trials being need to achieve conventional reliability.

Position and variability at the kinematic landmarks The results for the position at positive peak acceleration, peak velocity, negative peak acceleration, and movement termination are also shown in Table 2. The person-by-trial variance was one-half to two-thirds of the total variance for the first half of the movement. This resulted in eight and five trials being needed to achieve conventional reliability at positive peak acceleration and peak velocity. The percentage of person-by-trial variance increased

Table 1 G and D study results for the basic measurements of the movement

	G Study				D Study
	M ± *SE*	*Person* (%)	*Trial* (%)	*P* × *T* (%)	*g* ≥ .80
Reaction time					
Choice right long	284 ± 9.0 ms	39.8	5.1	55.1	6
Simple right long	262 ± 6.5 ms	28.6	6.0	65.4	10
Movement time	304 ± 12.2 ms	62.5	0.2	37.3	3
Constant error	1.94 ± 0.45 mm	7.9	0.0	92.1	47

Standard error (*SE*) was calculated by calculating the mean across all trials for each participant. The standard deviation of these means was calculated and then divided by the square root of the number of participants

Table 2 G and D study results for the time, magnitude, and position at the kinematic landmarks

		G Study			D Study
	M ± *SE*	*Person* (%)	*Trial* (%)	*P* × *T* (%)	*g* ≥ .80
Time (ms)					
Positive peak acceleration	41 ± 3.8	44.5	0.0	55.5	5
Peak velocity	134 ± 5.7	64.5	0.0	35.5	3
Negative peak acceleration	248 ± 10.8	56.52	1.44	42.04	3
Magnitude					
Positive peak acceleration (*mm/s*^{2})	14,673 ± 1,142	75.6	0.1	24.3	2
Peak velocity (*mm/s*)	1,174 ± 52	80.2	0.4	19.4	1
Negative peak acceleration (*mm/s*^{2})	– 12,930 ± 1,586	80.135	0.230	19.635	1
Position (mm)					
Positive peak acceleration	11 ± 1.3	34.0	0.0	66.0	8
Peak velocity	92 ± 1.9	44.42	0.55	55.03	5
Negative peak acceleration	187 ± 2.2	21.4	0.6	78.0	15
Movement termination	202 ± 0.5	7.9	0.0	92.1	47

in the second half of the movements. This resulted in 15 and 47 trials being needed to achieve conventional reliability at negative peak acceleration and movement termination. Note that the G and D study results for the position at movement termination and constant error (Table 1) are identical. Constant error is simply the position at the end of the movement minus the position of the middle of the target, and this subtraction does not change the statistical analysis.

Recall that the variability at kinematics landmarks is dependent on the position at the landmarks. We can, therefore, apply the results of the G and D studies on the positions at the kinematic landmarks to the variability. Thus, the same number of trials are needed for either reliable positions or variability at the kinematic landmarks. The largest number of trials (47) was required for the variability at movement termination, which is also called the variable error.

Time-normalized displacement profiles At 10% of the movement, the total variance was split between person (57.0%) and person-by-trial (42.7%; Table 3). A D study estimated that three trials were needed to achieve the conventional level of reliability. As the movement progressed, the percentage of person variance decreased, the person-by-trial variance increased, and the number of trials required for conventional reliability increased. There were larger increases in the person-by-trial variance and the number of trials required at the end of the movement. The number of trials to achieve the conventional level of reliability doubled from 80% to 90% (12 to 23), and then doubled again from 90% to 100% (23 to 47). Note, again, that the G and D study results for 100% of the movement, position at movement termination (Table 2), and constant error (Table 1) were identical.

We took a closer look at the time-normalized displacement profile results by plotting the absolute variance estimates attributed to person and to the person-by-trial interaction (instead of the percentage of total variance; Fig. 2). In both cases, the variance increased from 10% to 50% of the movement, and then decreased from 50% to 100%. There were two important differences between the variance for person and for the person-by-trial interaction. First, the person-by-trial interaction had visibly larger variance from 30% to 100%. Second, the person variance decreased to almost zero (3.0 mm^2) at 100%, whereas the variance of the interaction was 35.2 mm^2. The person-by-trial variance at 100% was smaller than earlier in the movement, but it caused the vast majority of the total variance. That is why the person-by-trial interaction caused 92.1% of the total variance at the end of the movement.

Correlation analysis The mean correlations between positive peak acceleration, peak velocity, negative peak acceleration, and movement termination are shown in Fig. 3; the

Table 3 G and D study results for the time-normalized displacement profiles

		G Study			D Study
	M ± *SE* (mm)	*Person* (%)	*Trial* (%)	*P* × *T* (%)	*g* ≥ .80
10%	5 ± 0.7	57.0	0.3	42.7	3
20%	20 ± 1.3	49.2	0.3	50.5	5
30%	44 ± 1.8	41.5	0.0	58.5	6
40%	75 ± 2.3	38.6	0.0	61.4	7
50%	108 ± 2.5	37.0	0.0	63.0	7
60%	139 ± 2.3	35.3	0.0	64.7	8
70%	165 ± 1.8	32.8	0.0	67.2	9
80%	184 ± 1.3	26.1	0.0	73.9	12
90%	197 ± 0.7	14.9	0.0	85.1	23
100%	202 ± 0.5	7.9	0.0	92.1	47

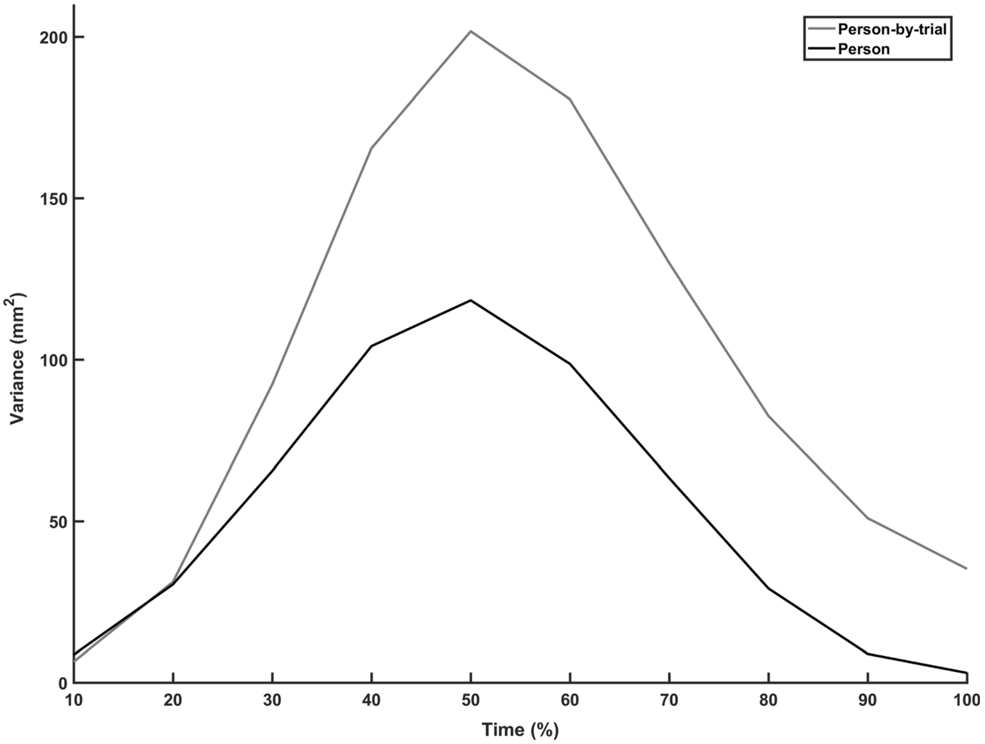

Fig. 2 Absolute variance estimates of the time-normalized displacement profiles attributed to person (black line) and the person-by-trial interaction (gray line)

correlations for all 32 trials are shown at the far right. They showed the typical pattern for movements with vision, which is a small coefficient of determination at positive peak acceleration, a medium coefficient of determination at peak velocity, and medium-to-large coefficient of determination at negative peak acceleration. The simulated correlations had a similar pattern for the three kinematic landmarks; the coefficients of determination were very large with three trials, and exhibited exponential decay toward the actual coefficients of determination with 32 trials as the number of trials in the subset increased.

The simulated correlations for each kinematic landmark were compared to the actual correlations with a series of 28 paired-samples *t* tests. For positive peak acceleration, the simulated correlation with 28 trials was not significantly different from the actual correlation. The simulated correlations were not significantly different for 24 and 27 trials for peak velocity. These results suggest that using fewer than 32 trials significantly decreases the accuracy of the coefficients of determination. As for negative peak acceleration, the simulated correlations were not significantly different for 15, 19, 23, 24, 26, and 28–30 trials. This suggests that using 28 trials instead of 32 should not significantly affect the accuracy of the coefficient of determination. It might be possible to use fewer trials, but the comparisons varied between significantly different and not significantly different between 15 and 27 trials.

Variability of the D study results in the eight conditions

The D study results were similar but not identical in the eight conditions (choice reaction time: right long, right short, left long, left short; simple reaction time: right long, right short, left long, left short). We already noted that for reaction time, fewer trials were required for choice reaction time movements (six, six, five, and three trials, respectively) than for simple reaction time movements (10, 15, 10, and 14 trials) to achieve the conventional level of reliability. Reaction time was the only dependent variable for which there was a clear difference in the number of trials between choice and simple reaction time tasks. There was a large amount of variability in the number of trials required for a reliable estimate of constant error, ranging from 30 trials for choice, right, short movements, to 89 trials for simple, right, short movements, and an outlier of 297 trials for simple, left, short movements. These results were identical for the position at movement termination and the time-normalized displacement at 100%, because they are all based on the same data (these repetitions were excluded from Table 4). Finally, there was a medium amount of variability for all of the position variables and most of the time-normalized displacement variables. The time-normalized displacement at 80%, for example, required 12 trials for choice, right, long or simple, left, short movements, and 24 trials for choice, left, short or simple, right, short movements. The results for all of the G and D studies and the data are available as supplementary material.

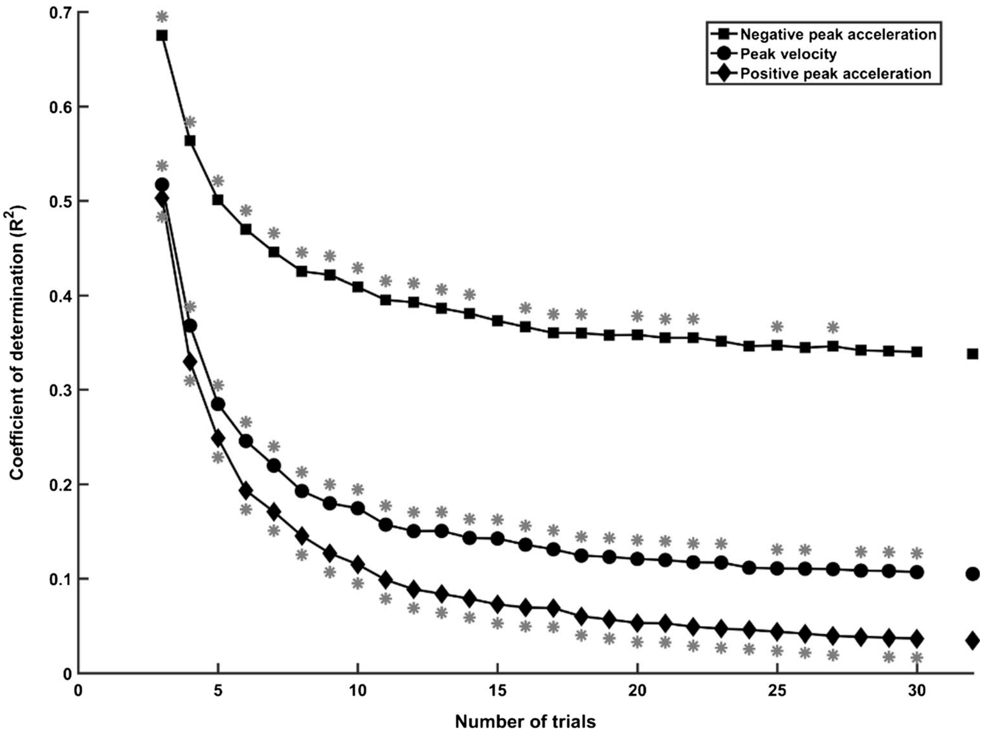

Fig. 3 Simulated and actual mean correlations between positive peak acceleration, peak velocity, negative peak acceleration and movement termination. The simulated correlations involved random subsets of 3 to 30 trials. The actual correlations involved all 32 trials. $^*p < .05$

Discussion

The purpose of the present analysis was to estimate the minimum number of trials per participant required to achieve the conventional level of reliability for trajectory analysis of discrete goal-directed pointing movements. We analyzed the basic measurements of movement (reaction time, movement time, constant error) and three common methods of trajectory analysis (kinematic landmarks, spatial variability, and time-normalized displacement profiles) within the framework of G theory. G studies were used to decompose the total variance of these variables into the percent contributions from person, trial, and the person-by-trial interaction. D studies were then used to estimate the minimum number of trials per participant required to achieve the conventional level of reliability.

The first novel finding was that more trials were needed for reliable reaction time data in simple reaction time tasks (ten trials per participant) than choice reaction time tasks (six trials). One might think that because simple reaction time is shorter and, typically, less variable than choice reaction time (Luce, 1986) that it would have less person-by-trial variance, but the opposite was true. This is an example of how statistical analysis with G theory is required to decompose the total variance into its sources. Movement time required three trials—fewer trials than required for reaction time—to achieve conventional reliability. Constant error, in contrast, required 47 trials. Recall that this is the number of good test trials per condition. If a researcher was using a four-choice reaction time test, then they would need 188 good test trials. That would be a total of 218 trials for the experiment, assuming 5% of the trials are bad (target misses, anticipation, etc.) and the experiment began with 20 practice trials. It is possible that some pointing studies that rely on constant error have enough trials for reliable data, but most probably do not. The experiment that these data came from (Blinch et al., 2014), for example, had 32 good test trials per condition, yielding a reliability coefficient of .73 for constant error.

The time and magnitude of all kinematic landmarks required five trials or fewer to achieve the conventional level of reliability. One might suspect that acceleration landmarks would require more trials than velocity landmarks. Differentiation increases the noise in a signal, and acceleration is the second derivative of displacement whereas velocity is the first derivative. Thus, acceleration has more noise than velocity. The acceleration and velocity landmarks required a similar number of trials despite this difference in noise.

The number of trials required for position and variability at the kinematics landmarks mostly increased throughout movement execution. Positive peak acceleration and peak velocity require eight and five trials to achieve the conventional level of reliability. This increases to 15 trials for negative peak acceleration and then 47 trials for movement termination. Variability analysis typically involves all these kinematic landmarks, and so the number of trials required will depend on whichever landmark requires the most trials, which in this case was movement termination. Variability analysis should,

Table 4 Descriptive statistics for the D study results (the minimum number of trials required to achieve the conventional level of reliability) in the eight conditions

	M ± *SD*	Minimum	Maximum
Reaction time	8.6 ± 4.3	3	15
Movement time	2.9 ± 1.1	2	5
Constant error	90.9 ± 85.8	30	297
Time			
Positive peak acceleration	5.4 ± 1.8	3	8
Peak velocity	3.5 ± 1.1	2	5
Negative peak acceleration	3.3 ± 1.0	2	5
Magnitude			
Positive peak acceleration	2.0 ± 0	2	2
Peak velocity	1.6 ± 0.5	1	2
Negative peak acceleration	2.0 ± 0.5	1	3
Position			
Positive peak acceleration	9.0 ± 3.1	4	14
Peak velocity	12.4 ± 8.9	5	31
Negative peak acceleration	14.4 ± 5.8	8	27
Time-normalized displacement			
10%	5.5 ± 2.2	3	10
20%	5.0 ± 0.9	4	6
30%	5.1 ± 1.4	3	7
40%	6.0 ± 1.9	3	9
50%	7.0 ± 2.0	4	11
60%	8.4 ± 2.5	5	13
70%	10.8 ± 2.7	8	15
80%	17.4 ± 5.2	12	24
90%	39.5 ± 14.4	23	63

therefore, include at least 47 trials. This will achieve the conventional level of reliability at movement termination and an even higher level of reliability at the other kinematic landmarks. One limitation with the present data is that we tested only movements with full visual feedback. Spatial variability analysis typically involves comparing movements with full visual feedback to movements with reduced or no visual feedback. Those types of movements should be tested to determine whether the number of trials required to achieve the conventional level of reliability differs.

As for the time-normalized displacement profiles, the number of trials to achieve conventional reliability slowly increased from three trials at 10%, to 12 trials at 80%. The number of trials then doubled from 80% to 90%, and then doubled again from 90% to 100% (23 trials at 90% and 47 trials at 100%). The absolute variance showed that the person and person-by-trial variance increased from 10% to 50% and then decreased from 50% to 100%. This pattern of increasing and then decreasing variability is indicative of online corrections in the latter half of the movements (Khan et al., 2006), which allow participants to hit the targets as quickly and accurately as possible (Carlton, 1994). Interestingly, the online corrections eliminated the person true variance but could only reduce the larger person-by-trial random error variance. This caused the large increase in the number of trials required to achieve conventional reliability at 90% and 100% of the movement. So, although the online corrections were beneficial for the movement accuracy, they had the detrimental effect of drastically increasing the number of trials required for minimum reliability.

It would be interesting to compare the reliability of movements with online corrections to movements with reduced online control (e.g., ballistic movements, movements without visual feedback, or movements with a temporal goal). We predict that the increase in variance from 10% to 50% would continue or plateau for movements with less online control. The number of trials required for conventional reliability would likely slowly increase, preventing the large increase at the end of the movements. If this is the case, then fewer trials might be needed for reliable time-normalized displacement profiles with reduced online control.

An important issue is the generalizability of these results. Recall that G and D studies generalize to movements with similar measurement conditions. We examined unimanual pointing movements with the typical instructions to hit the targets as quickly and accurately as possible. Other types of pointing movements with the same instructions should have similar G and D study results—for example, movements with similar amplitudes and target widths. In fact, the present analysis found comparable results for long (20 cm) and short (10 cm) movement amplitudes. All the results were similar for movements in simple and two-choice reaction time tasks, save for the minimum number of trials required to achieve conventional reliability for reaction time. The results for the other basic measurements of movement and the trajectory analysis should generalize to other choice reaction time tasks. It is also likely that our results for pointing movements will generalize to the transport component of reach-to-grasp movements. Dissimilar movements with different instructions should be tested to determine their G and D study results. Common examples are ballistics movements made as quickly as possible, movements without visual feedback (occluded vision or a proprioceptive target), movements with a temporal goal, and movements with online corrections to target jumps or perturbations.

The correlation analysis could not be analyzed with G theory and so we used a simulation analysis instead. The results suggest that using fewer than 32 trials significantly decreases the accuracy of the coefficients of determination for positive peak acceleration and peak velocity. Using 28 trials instead of 32 did not affect the accuracy of the coefficient of determination at negative peak acceleration. It might be possible to use even fewer trials, but some of the coefficients between 15 and 27 trials were less accurate than 32 trials. Correlation analysis typically involves analyzing all three kinematic landmarks. The number of trials required will, therefore, depend on whichever kinematic landmark requires the most trials, which were positive peak acceleration and peak velocity. For this dataset, we recommend the correlation analysis use all 32 trials.

Another way to interpret the correlation data is to determine when the difference between the simulated and actual coefficients of determination is less than a practically significant amount. Although some differences still might be significant at that point, these differences might be too small to be important. We reviewed several articles with correlation analysis to estimate the smallest differences in the coefficients of determination that were significantly different and practically significant (de Grosbois & Tremblay, 2016; Heath, 2005; Heath et al., 2004; Khan & Franks, 2003; Krigolson & Heath, 2004). This different was .06, and so we checked when the differences between our stimulated and actual coefficients were less than half of that, .03. This occurred with 16 trials for negative peak acceleration, 17 trials for peak velocity, and 18 trials for positive peak acceleration. Therefore, if we base the number of trials required for correlation analysis on practical significant, then 18 trials may be sufficient.

This simulation analysis is, unfortunately, limited by the number of trials we collected, and so we cannot extrapolate to what might happen with more trials. D studies have the advantage that they can extrapolate beyond the number of trials collected. That is why the D study on constant error suggested that more than 32 trials (47) were required to achieve the conventional level of reliability. Another limitation with the present data, which we already mentioned concerning the spatial variability analysis, is that we tested only movements with full visual feedback. Movements with reduced or no visual feedback should be tested to determine whether the accuracy of the coefficients of determination are different than movements with full visual feedback.

Conclusions

The number of trials per participant needed for reliable data of discrete goal-directed pointing movements depended on the dependent variable. Reaction time required ten and six trials per participant to achieve the conventional level of reliability in simple and choice reaction time tasks, respectively. Movement time required three trials, constant error required 47, and the time and magnitude of the kinematic landmarks required five trials or fewer. Position and variability at all the kinematic landmarks required 47 trials to meet or exceed the conventional level of reliability. The time-normalized displacement profiles required 12 trials or fewer from 10% to 80% of the movement. This increased to 47 trials at 100% of the movement, which was the same as constant error. Finally, all 32 trials were required for the correlation analysis to prevent reducing the accuracy of any of the coefficients of determination. The correlation analysis, however, could probably be reduced to 18 trials without a practical decrease in accuracy.

One thing is certain when considering measurement reliability: Collecting more trials will increase the reliability. This, however, needs to be offset by the cost to researchers and participants of collecting an enormous number of trials. Our results can be used to inform future research and ensure reliable measurement of discrete pointing movements with the lowest cost, which is an important aspect of improving experimental quality. When applying these results, you will need to consider all the dependent variables that you wish to measure. For example, if you are interested in choice reaction time (six trials for conventional reliability) and variability analysis (47), you should ensure that there are at least 47 good trials in each condition. That should be enough trials to achieve the conventional level of reliability for the variability analysis and more than enough trials for reaction time. Importantly, these results should also apply to movements with similar measurement conditions. Additional experiments could estimate how many trials are needed for the reliable measurement of dissimilar movements.

Author note All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or the national research committee, and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. Our thanks to Tim N. Welsh and two anonymous reviewers for their insightful critiques. The authors declare that they have no conflicts of interest.

References

Bernstein, N. A. (1967). *The coordination and regulation of movements*. Oxford: Pergamon Press.

Blinch, J., Cameron, B. D., Cressman, E. K., Franks, I. M., Carpenter, M. G., & Chua, R. (2014). Comparing movement preparation of unimanual, bimanual symmetric, and bimanual asymmetric movements. *Experimental Brain Research*, *232*, 947–955.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Broglio, S. P., Zhu, W., Sopiarz, K., & Park, Y. (2009). Generalizability theory analysis of balance error scoring system reliability in health young adults. *Journal of Athletic Training*, *44*, 497–502.

Carlton, L. G. (1994). The effects of temporal-precision and time-minimization constraints on the spatial and temporal accuracy of aimed hand movements. *Journal of Motor Behavior*, *26*, 43–50.

Carlton, M., J., Newell, K. M., & Carlton, L. G. (1984). Predicting individual discrete response outcomes from kinematic characteristics. Journal of Human Movement Studies, 10, 62-82.

Crocker, P. R. E., Bailey, D. A., Faulkner, R. A., Kowalski, K. C., & McGrath, R. (1997). Measuring general levels of physical activity: Preliminary evidence for the Physical Activity Questionnaire for Older Children. *Medicine and Science in Sports and Exercise*, *29*, 1344–1349.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.

de Grosbois, J., & Tremblay, L. (2016). Quantifying online visuomotor feedback utilization in the frequency domain. *Behavior Research Methods*, *48*, 1653–1666.

Desmurget, M., Pélisson, D., Rossetti, Y., & Prablanc, C. (1998). From eye to hand: Planning goal-directed movements. *Neuroscience & Biobehavioral Reviews*, *22*, 761–788.

Doyle, R. J., Hsiao-Wecksler, E. T., Ragan, B. G., & Rosengren, K. S. (2007). Generalizability of center of pressure measures of quiet standing. *Gait and Posture*, *25*, 166–171.

Elliott, D., Helsen, W. F., & Chua, R. (2001). A century later: Woodworth's (1899) two-component model of goal-directed aiming. *Psychological Bulletin*, *127*, 342–357.https://doi.org/10.1037/0033-2909.127.3.342

Elliott, D., Lyons, J., Hayes, S. J., Burkitt, J. J., Roberts, J. W., Grierson, L. E., … Bennett, S. J. (2017). The multiple process model of goal-directed reaching revisited. *Neuroscience & Biobehavioral Reviews*, *72*, 95–110.

Engelsen, B. K., & Hagtvet, K. A. (1999). A generalizability study of the Eating Attitudes Test. (EAT-12). in non-clinical adolescents. *Eating and Weight Disorders*, *4*, 179–186.

Gaveau, V., Pisella, L., Priot, A. E., Fukui, T., Rossetti, Y., Pélisson, D., & Prablanc, C. (2014). Automatic online control of motor adjustments in reaching and grasping. *Neuropsychologia*, *55*, 25–40.

Gordon, J., & Ghez, C. (1987). Trajectory control in targeted force impulses. III. Compensatory adjustments for initial errors. *Experimental Brain Research*, *67*, 253–269.

Hansen, S., Elliott, D., & Khan, M. A. (2008). Quantifying the variability of three-dimensional aiming movements using ellipsoids. *Motor Control*, *12*, 241–251.

Heath, M. (2005). Role of limb and target vision in the online control of memory-guided reaches. *Motor Control*, *9*, 281–311.

Heath, M., Westwood, D. A., & Binsted, G. (2004). The control of memory-guided reaching movements in peripersonal space. *Motor Control*, *8*, 76–106.

Jeannerod, M. (1984). The timing of natural prehension movements. *Journal of Motor Behavior*, *16*, 235–254.

Kang, M., Bjornson, K., Barreira, T. V., Ragan, B. G., & Song, K. (2014). The minimum number of days required to establish reliable physical activity estimates in children aged 2–15 years. *Physiological Measurement*, *35*, 2229–2237.

Khan, M. A., & Franks, I. M. (2003). Online versus offline processing of visual feedback in the production of component submovements. *Journal of Motor Behavior*, *35*, 285–295.

Khan, M. A., Franks, I. M., Elliott, D., Lawrence, G. P., Chua, R., Bernier, P. M., … Weeks, D. J. (2006). Inferring online and offline processing of visual feedback in target-directed movements from kinematic data. *Neuroscience & Biobehavioral Reviews*, *30*, 1106–1121.https://doi.org/10.1016/j.neubiorev.2006.05.002

Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adaptive Physical Activity Quarterly*, *29*, 346–365.

Krigolson, O., & Heath, M. (2004). Background visual cues and memory-guided reaching. *Human Movement Science*, *23*, 861–877.

Llabre, M. M., Ironson, G. H., Spitzer, S. B., Gellman, M. D., Weidler, D. J., & Schneiderman, N. (1988). How many blood-pressure measurements are enough: An application of generalizability theory to the study of blood-pressure reliability. *Psychophysiology*, *25*, 97–106.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

MacKenzie, C. L., Marteniuk, R. G., Dugas, C., Liske, D., & Eickmeier, B. (1987). Three-dimensional movement trajectories in Fitts' task: Implications for control. *Quarterly Journal of Experimental Psychology*, *39*, 629–647.

Marteniuk, R. G., MacKenzi, C. L., Jeannerod, M., Athenes, S., & Dugas, C. (1987). Constraints of human arm movement trajectories. *Canadian Journal of Psychology*, *41*, 365–378.

Melmoth, D. R., Storoni, M., Todd, G., Finlay, A. L., & Grant, S. (2007). Dissociation between vergence and binocular disparity cues in the control of prehension. *Experimental Brain Research*, *183*, 283–298.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*, *38*, 542–547.

Oostwound Wijdenes, L., Brenner, E., & Smeets, J. B. (2014). Analysis of methods to determine the latency of online movement adjustments. *Behavior Research Methods*, *46*, 131–139.

Prablanc, C., Desmurget, M., & Gréa, H. (2003). Neural control of online guidance of hand reaching movements. *Progress in Brain Research*, *142*, 155–170.

Roebroeck, M. E., Harlaar, J., & Lankhorst, G. T. (1993). The application of generalizability theory to reliability assessment: An illustration using isometric force measurements. *Physical Therapy*, *73*, 386–395.

Schot, W. D., Brenner, E., & Smeets, J. B. (2010). Robust movement segmentation by combining multiple sources of information. *Journal of Neuroscience Methods 187*, 147–155.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*, 301–317.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (in press). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*https://doi.org/10.1037/met0000107

Welk, G. J., Schaben, J. A., & Morrow, J. R., Jr. (2004). Reliability of accelerometry-based activity monitors: a generalizability study. *Medicine and Science in Sports and Exercise*, *36*, 1637–1645.

Welsh, T. N., Higgins, L., & Elliott, D. (2007). Are there age-related differences in learning to optimize speed, accuracy, and energy expenditure? *Human Movement Science*, *26*, 892–912.

Winter, D. A., Sidwall, H. G., & Hobson, D. A. (1974). Measurement and reduction of noise in kinematics of locomotion. *Journal of Biomechanics*, *7*, 157–159.

Woodworth, R. S. (1899). The accuracy of voluntary movement. *Psychological Review*, *3*(Suppl. 13), 1–114.https://doi.org/10.1037/h0092992