Available online at www.sciencedirect.com

ScienceDirect

Journal of Sport and Health Science 8 (2019) 527−531

www.jshs.org.cn

Opinion

Variability: Human nature and its impact on measurement and statistical analysis

Heng Li [a], Zezhao Chen [b], Weimo Zhu [b,*]

[a] *Department of Physical Education, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China*

[b] *Department of Kinesiology & Community Health, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

Received 21 January 2019; revised 3 February 2019; accepted 5 June 2019
Available online 18 June 2019

The world is colorful, different, and diverse! So are humans. Human variability refers to the variability among individuals, which could be the variability of a human trait (e.g., body fat percent), or the difference in the response of a trait to a simulation (e.g., losing or not losing weight when facing the same intervention). In research, the most commonly studied types of variability are between-individual variability and within-individual variability. Between-individual variability is the difference among individuals (e.g., the differences in height among individuals). Within-individual variability refers to the variability of an individual at different times (e.g., the difference of one's weight, performance, and mood at different times). Variability is also a common phenomenon in human performance.[1,2]

Is a large variability bad? Should an outlier in a data set be considered as a part of variability? What is the impact of variability on commonly used measurement and statistical methods? Variability has long been of interest in human-movement research. A set of measurement (e.g., reliability coefficients) and statistical indexes (e.g., standard deviation (SD) and variance) have been developed to measure and analyze variability. However, due to the complex nature of variability, and the lack of advanced measurement techniques and statistical training for researchers, misunderstandings of variability often occur. As a result, variability in research has often been analyzed incorrectly and has led to findings being interpreted erroneously. Here, we summarize common errors related to variability and how to address them. We hope that this discussion helps researchers to understand variability better, and thus contributes to its proper use.

1. Common errors in measuring variability

A common error in measuring variability is ignoring the sensitivity of the measurement. Sensitivity, in this context, is defined as the ability to discriminate differences. In practice, sensitivity is the ability to measure variability in stimuli or responses, detect a change, or classify a status. Without appropriate sensitivity, a difference, a change, or a different status may not be detected. For example, if meaningful changes in a child's height are in centimeter (cm) units, but the test administrators use a ruler with the smallest unit in inches (2.54 cm), the measurement tool may not have the needed sensitivity. However, greater sensitivity may not always be better. For example, using a ruler with millimeter (mm) units to measure height may not be appropriate, since such a small mm change in height may be associated with natural fluctuations within a day, and thus may provide a sense of accuracy that is not present. Therefore, it is important to understand the degree of variability to be measured, what a meaningful variability is, and whether the measure is sensitive enough to detect meaningful variability.

Another "error", also related to the measures used, involves mixing the variability of humans and measures. For example, researchers have reported the "reliability" of physical activity measurement devices using the following design: Ask a group of subjects to wear a device for 3 days, 7 days, or more days and use the data collected to make conclusions about the reliability of the device. Obviously, this analysis cannot be used to evaluate the day-by-day variability of the measurement device, because variability in daily subjects' physical activity behavior is (likely a big) part of the variability that is measured. To measure the reliability of a device, a different repeated-measurement design should be used (e.g., ask participants to repeat their walking for the same distance, or exercising for the same duration in the same environment at a single point in time[3]). To distinguish different types of variability, there has been a call for eliminating the term "reliability", and instead replace it by terms such as "score reliability" (when all variabilities are mixed together), "personal stability" (when measuring intra-individual variability), and "instrument reliability" (when measuring intra-instrument variability). The last type of variability

Peer review under responsibility of Shanghai University of Sport.

* Corresponding author.

E-mail address: weimozhu@uiuc.edu (W. Zhu).

https://doi.org/10.1016/j.jshs.2019.06.002

Table 1
Data for examples 1, 2, and 3.

Examples											
Data 1											
	ID	Pre-test	Post-test								
	1	3	30								
	2	5	40								
	3	7	70								
	4	9	90								
	5	11	110								
Data 2											
	ID	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
	1	101	101	101	100	100	100	101	101	101	101
	2	99	100	96	99	99	101	99	101	100	99
	3	101	100	100	89	100	102	101	101	102	100
	4	100	104	100	102	101	102	100	102	101	99
	5	99	98	100	100	100	100	100	101	101	100
Data 3											
ID	1	2	3	4	5	6	7	8	9	**10**	11
1.5-mile run (min)	12.2	15.0	11.5	13.1	12.5	14.2	15.2	13.5	11.34	**19.0**	14.11
VO_{2max} (mL/kg/min)	49	43	52	41	47	45	41	49	51	**51**	48

Note: While Subject 10 had a poor performance in 1.5-mile run, his/her VO_{2max} was the second best in the data set, which made this pair data "outlier".
Abbreviations: T = trial; VO_{2max} = maximal oxygen uptake.

can be further broken down into "location invariance" (when variability in the location of where the device is worn is being investigated) and "device equivalence" (when between-device variability is being studied[3,4]).

Failing to recognize the potential impact of variability on measurement coefficients is another common error. Using the Pearson correlation coefficient as a measure of reliability may illustrate this point. The Data 1 in Table 1 represent a hypothetical test−retest data set. By glancing at the data, one can easily detect that the test and retest are not consistent. Furthermore, it can be seen that the inconsistency is systematic because higher pretest scores seem to be associated with larger differences between the test−retest scores. Using the Pearson correlation coefficient for this data set gives a result of $r = 0.99$, an almost perfect correlation! However, does this strong correlation also mean a strong reliability? The answer, of course, is no! The incorrect estimation of the reliability or variability is due to the fact that the Pearson correlation is biased by the order of 2 data sets. As long as the order of a set of a pair of data is kept the same or similar, the correlation will be high, even if there is a large absolute difference between the pairs. This limitation of the Pearson correlation coefficient can be overcome by applying a regression analysis in which both slope and intercept are examined simultaneously: A slope of 1.0 or near 1.0 and an intercept of 0.0 or near 0.0 indicate a high test−retest reliability; a slope of 1.0 or near 1.0, but an intercept far from 0.0 indicate a poor test−retest reliability caused likely by a systematic error; and finally, a slope far from 1.0 and an intercept far from 0.0 indicate a poor test−retest reliability. Another commonly used approach to overcome the limitations is to use an interclass coefficient (ICC) calculation. The relationship among reliability (R), variability, and ICC can be explained using Eq. (1), in which reliability is defined as the ratio of the variability between subjects' true scores (VT), and variability between subjects' obtained scores (VB), which includes VT and an error:

$$\text{Reliability (R)} = \frac{\text{Variability between subjects' true scores (VT)}}{\text{Variability between subjects' obtained scores (VB)}} = \frac{\text{VT}}{\text{VT} + \text{Error}} \qquad \text{Eq. (1)}$$

According to Eq. (1), when there is no error (error = 0), reliability is perfect (=1). In contrast, when everything observed is an error, VT becomes 0, and reliability will be equal to 0, too. VT can be considered the variability among subjects, which is expressed as $MS_{between} - MS_{within}$ in the context of a two-way analysis of variance (ANOVA; see Refs 5 and 6 for a variety of ICCs and their applications), whereas VB can be considered as subject variability plus error, which can be represented by $MS_{between}$ in ANOVA testing. In ANOVA terms, Eq. (1) becomes:

$$\text{R} = \frac{\text{Subject variability}}{\text{Subject variability} + \text{Error}} = \frac{MS_{between} - MS_{within}}{MS_{between}} \qquad \text{Eq. (2)}$$

By applying Eq. (2) to Data 1, the systematic error is detected and taken into consideration, and the new reliability coefficient becomes 0.31:

$$\text{R, or ICC} = \frac{670 - 460}{670} = 0.31 \qquad \text{Eq. (2a)}$$

However, the ICC does not take care of all reliability problems caused by variability. Consider Data 2 in Table 1, which is a small sample from a real study in which the reliability of a pedometer instrument was evaluated.[3] Specifically, subjects were asked to wear 10 pedometers and walk 100 steps 10 times in a row. Data 2 is a sample from 5 subjects. In contrast with Data 1, the variability among trials was small in this experiment and most results were close to the correct value of 100. Using Eq. (2) for this data set, one obtains a low ICC coefficient: 0.34!

$$\text{R, or ICC} = \frac{5.720-3.776}{5.720} = 0.34 \qquad \text{Eq. (2b)}$$

What went wrong? Again, variability is the problem! But in this case, it is the small variability. More specifically, it was due to the fact that everyone was asked to walk the same 100 steps, so the small between-subject variability among trials caused the problem. As a result, the within-subjects and between-subjects variabilities became similar, so R, or the ratio in Eq. (2b), became small. As illustrated in Table 2, Pearson correlations also failed this time due to the lack of variability, and most of the computed between-trial correlations were low. Thus, the pedometers were so reliable (or the variability between trials was so small) that they caused 2 commonly used reliability coefficients to fail! These 2 opposite variability impacts, one from the large variability and the other from the small one, indicate that when applying measurement coefficients, the degree of variability for all variables, as well as their potential impact on a specific coefficient, should be carefully examined.

Another common error in human performance research involves failing to understand the measures of variability, and applying them incorrectly. Table 3 summarizes a set of commonly used variability measures, including their advantages

Table 2
Correlations among trials (Ts).

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	–									
T2	0.34	—								
T3	0.64	0.16	—							
T4	−0.49	0.31	−0.05	—						
T5	0.35	0.65	0.73	0.21	—					
T6	0.25	0.57	−0.13	−0.44	0.35	—				
T7	0.90	0.06	0.86	−0.46	0.42	0.00	—			
T8	0.00	0.87	0.17	0.43	0.79	0.56	−0.13	—		
T9	0.71	0.00	0.73	−0.69	0.50	0.35	0.85	0.00	—	
T10	0.60	−0.33	0.67	−0.17	0.00	−0.60	0.79	−0.54	0.42	–

Table 3
Commonly used measures of variability.

Term	Mathematical definition	Advantages	Limitations
Range (R)	R = Highest score − Lowest score	Easy to use; a quick assessment on the overall variability of the data.	Determined by only 2 numbers in a data set and easily biased by an outlier.
Sample variance (S)	$S^2 = \frac{\Sigma(X–M)^2}{N–1}$, where X = raw score, M = mean, and N = sample size	One of the commonly used variability indexes, in which the variability is computed based the average of the difference of each data point in the sample from their mean.	Because it is derived from the squared differences, it is not original unit.
Sample standard deviation (SD)	$S(SD) = \sqrt{\frac{\Sigma(X–M)^2}{N–1}}$	SD is simply the square root of variance. When the data are normally distributed, SD, together with the mean, provides the degree of the variability related to the percent of the distribution covered: ±1 SD covers 68% of the distribution, ±2 SD covers 95%, and ±3 SD covers 99%.	When the distribution is skewed, however, the mean and SD are not the best measures for the central tendency and variability. Instead, median and interpercentile range should be used. The most commonly used measure is interquartile range (IQR).
Interquartile range (IQR), the most commonly used interpercentile	IQR = $X_{0.75} - X_{0.25}$, where $X_{0.75}$ = 75th percentile and $X_{0.25}$ = 25th percentile	The IQR is a measure of variability based on dividing a data set into quartiles, specifically the variability between the middle 50% of the distribution and in variance of the shape of the distribution. IQR is a useful variability measure when there are extreme outliers.	The IQR is not amenable to mathematical manipulation.
Coefficient of variation (CV) or relative SD	CV = SD/M × 100%	A simple way to determine the extent of variability in relation to the mean of the sample.	The CV may be misleading when there are negative values in the data or when the mean is close to zero.
Root mean square difference (RMSD)	$\text{RMSD} = \sqrt{\frac{\Sigma(\text{Measured values} - \text{Estimated values})^2}{N}}$	A variability measure that reflects the average of the squared difference between measured values and estimated values.	RMSD provides no information on the direction of incorrect estimation (i.e., overestimation or underestimation).
Mean signed difference (MSD)	$\text{MSD} = \frac{\Sigma(\text{Measured values} - \text{Estimated values})^2}{N}$	A simple variability measure for the direction of misestimating.	MSD provides limited information on the degree of misestimating.
Standard error of measurement (SE_{meas})	$SE_{meas} = SD\sqrt{(1-R)}$, where R = reliability of the measure	An SD-like estimate in classic testing theory for the variability range of an individual estimate.	Because the SE_{meas} is derived from SD and R, it may not provide an accurate estimate for scores in a specific region in the measure, e.g., for very small or large scores.

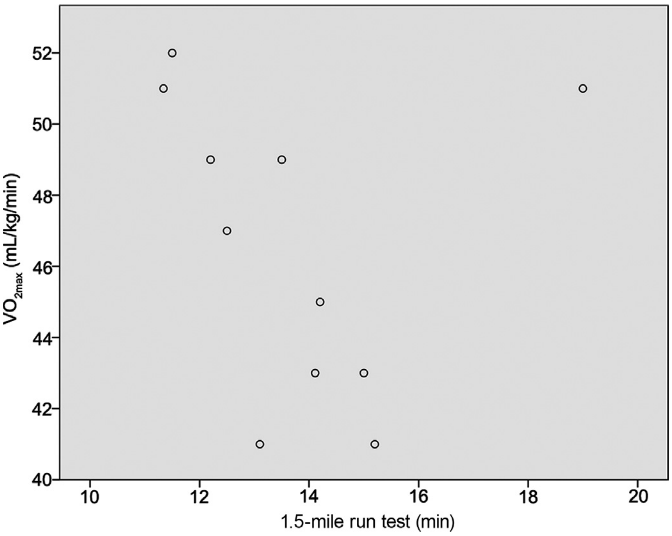

Fig. 1. Using a scatter plot to help identify the outlier. VO_{2max} = maximal oxygen uptake.

and limitations. The SD is probably the most commonly used variability measure. However, SD is sometimes applied to a skewed data distribution where the interpercentile range should be used instead. If, as described, the point of interest is to understand the impact of a variety of variabilities in measurement practice, such as person stability, instrument reliability, location invariance, and device equivalence, the method of generalizability theory is the most appropriate. Specifically, in generalizability theory, variance or variability is broken down using a carefully designed study and ANOVA-based analysis. Unfortunately, only few studies in human performance have taken advantage of this powerful approach, even though generalizability theory was introduced in the physical education literature[7,8] more than 40 years ago.

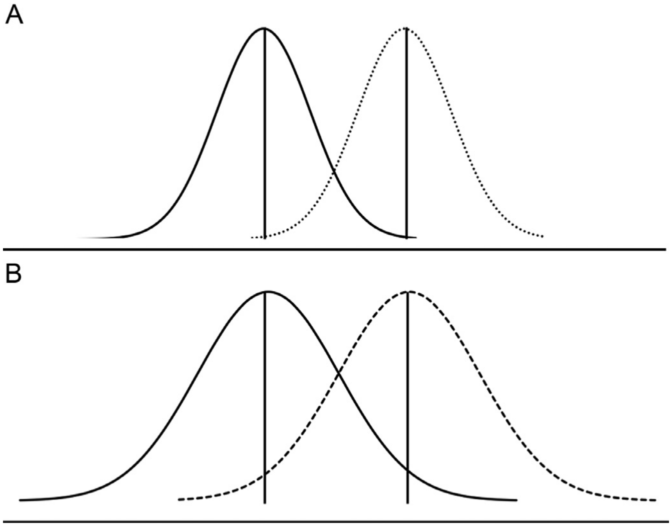

Fig. 2. Same treatment effect but with different variabilities (A, smaller; B, larger) in the control (*left*) and treatment (*right*) groups.

2. Common errors caused by variability when analyzing statistical data

As is the case of measuring variability, failure to recognize the impact of variability when analyzing data can also lead to errors. For a small data set, a single outlier may lead to a false conclusion. Let us use Data 3 in Table 1, another small data set randomly selected from a real study, in which the researchers were interested in determining if a 1.5-mile running time is valid to predict maximal oxygen uptake (VO_{2max}). To examine the validity of the 1.5-mile running time to predict VO_{2max}, the Pearson correlation was used, resulting in a coefficient of $r = -0.17$. Based on this short correlation, one might conclude that the 1.5-mile running time is not a valid predictor of VO_{2max}. Although the negative correlation indicates (correctly) that a low running time is associated with a high VO_{2max}, the correlation seems too low. Inspecting the data (Fig. 1), we detected an outlier (Subject 10). This subject had one of the highest VO_{2max} scores, but the slowest running time. If we were able to contact this subject, we could ask what had occurred during the running test and make a decision as to whether a retest was warranted. From a data analysis standpoint, we judge Subject 10 to be a clear outlier. Removing Subject 10 from analysis gives a Pearson correlation of $r = -0.82$. Based on this analysis, we reach the conclusion that the 1.5-mile running time may indeed be a valid measure for predicting VO_{2max}.

The impact of variability on parametric statistical analysis or null hypothesis testing can also be significant. Recall that the probability for rejecting a null hypothesis when it is false (or for detecting a difference when the treatment really works) is called "(statistical) power". There are 4 factors that affect power: the α (type I error) level, one- or two-tailed test, sample size, and effect size (ES). In practice, the α level is commonly set at 0.05 or 0.01. Also, most studies use a two-tailed testing approach. Regarding sample size, the chance of detecting a true difference becomes greater as sample size increases. Note, however, that a large sample size may result in the rejection of a null hypothesis whether there is a true treatment effect even if the difference between the treatment and control groups is small and does not have a practical meaning.[9,10] The last factor to affect the power of a statistic is the ES: the larger the ES, the higher the power. ES can be expressed using Eq. (3):[11]

$$ES = \frac{M_{treatment} - M_{control}}{SD_{pooled}} \qquad \text{Eq. (3)}$$

where $M_{treatment}$ = the mean of the treatment group, $M_{control}$ = the mean of the control group, and SD_{pooled} = pooled SD of the treatment and control groups. From Eq. (3), we see that the ES becomes large when the treatment effect is strong and variability in the treatment and control groups is small. For a given treatment effect, ES increases when variability decreases, and therefore, it is easier to obtain $p < 0.05$ and reject the null hypothesis (Fig. 2).

3. Conclusion

Variability is a natural part of the human condition and human performance. Understanding variability in all its forms, and selecting appropriate measurement techniques and statistical methods to best measure and analyze variability, is essential in scientific research related to human performance. Otherwise, the measurement tools used may fail to detect the true variability of human performance and led commonly used measurement and statistical indexes useless or misleading and research finding interpreted erroneously. Improving measurement techniques and providing better statistical training at the graduate level is thus urgently needed in human performance research.

Authors' contributions

HL carried out the idea and assisted the initial manuscript preparation; ZC made contributions to the early draft of this manuscript; and WZ involved in the study idea and design, final manuscript preparation, and modifications based on the feedback of reviewers and editors. All authors have read and approved the final version of the manuscript, and agree with the order of presentation of the authors.

Competing interests

The authors declare that they have no competing interests.

References

1. Davids K, Bennett S, Newell K, editors. *Movement system variability*. Champaign, IL: Human Kinetics; 2006.
2. Smith TJ, Henning R, Wade MG, Fisher T. *Variability in human performance*. Boca Raton, FL: CRC Press; 2015.
3. Zhu W, Lee M. Invariance of wearing location of Omron-BI pedometers: a validation study. *J Phys Act Health* 2010;**7**:706–17.
4. Zhu W. Reliability: what type, please! *J Sport Health Sci* 2013;**2**:62–4.
5. Baumgartner TA. Norm-referenced measurement: reliability. In: Safrit MJ, Woods TM, editors. *Measurement concepts in physical education and exercise science*. Champaign, IL: Human Kinetics; 1989.p.45–72.
6. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiroprac Med* 2016;**15**: 155–63.
7. Safrit MJ, Atwater AE, Baumgartner TA, West C, editors. *Reliability theory*. Washington, DC: American Alliance for Health, Physical Education and Recreation; 1976.
8. Morrow Jr JR. Generalizability theory. In: Safrit MJ, Woods TM, editors. *Measurement concepts in physical education and exercise science*. Champaign, IL: Human Kinetics; 1989.p.73–96.
9. Zhu W. Sadly, the earth is still round ($p < 0.05$). *J Sport Health Sci* 2012;**1**:9–11.
10. Zhu W. $p < 0.05$, < 0.01, < 0.001, < 0.0001, <0.0000, < 0.000001, or < 0.0000001. *J Sport Health Sci* 2016;**5**:77–9.
11. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.