Behav Res
DOI 10.3758/s13428-013-0318-1

A measure of within-participant response consistency

Justin A. MacDonald · David Trafimow

Abstract In this article, we introduce a measure of within-participant response consistency for use in the analysis of performance in decision-making tasks. The measure is an estimate of the correlation between the responses associated with two identical blocks of trials, the second of which has yet to be conducted. We derive a formula for the measure that can be applied to data from any two-choice decision task, including yes/no detection and two-alternative forced choice (2AFC). The estimate is easily calculated from the observed frequencies of hits, misses, false alarms, and correct rejections. We utilized data from an actual 2AFC experiment to compare estimated and observed consistency values; the estimates accounted for more than 90 % of the variability in observed consistency scores. We also discuss potential applications of the measure.

Keywords Consistency · Reliability · Decision making

The topic of measurement error has occupied psychologists for well over a century (e.g., Spearman, 1904). Not surprisingly, much effort has been devoted to addressing it by devising various ways to index reliability. One way to understand these different indexes is to think about the sources of error that they handle well or less well. Cronbach (1947; see also Schmidt, Le, & Ilies, 2003) has identified three such sources of error: random response, transient, and specific factor errors.

Random response errors occur when there are momentary variations in attention, distractions, and so on; these are errors across items. Transient errors are longitudinal and concern variations in mood, processing efficiency, learning, and so on across occasions. In contrast to transient errors, specific factor errors are consistent across occasions. An example would be that a particular item might be worded in such a way that it is interpreted differently than other items. Suppose that participants respond consistently to the wording of a particular item, either via exposure to the same item twice in a single occasion or across occasions, but differently to that item than to other items measuring the same construct. In this case, there would be no random response or transient error, but there would be a specific factor error. There would be variation in responses across items measuring the same construct. As Schmidt et al. (2003) explained, measures of internal consistency, of which the most prominent is Cronbach's alpha, address random response and specific factor errors, but they do not address transient error. In contrast, tests across occasions address transient error but do not address specific factor error. The measurement of transient errors requires estimating the variability of time series data collected from a single participant, which we will refer to as *within-participant variability* (or inversely, *consistency*). The most straightforward quantification of within-participant consistency is the variance of a participant's responses to a single stimulus across occasions. This measure is often referred to in the literature as *intraindividual variability* (see Nesselroade & Ram, 2004) and is affected by transient and random response errors but is unaffected by specific factor errors. There are several drawbacks to this approach, including the necessity of collecting responses to the same stimulus over multiple occasions and the fact that the consistency measure is based on responses to a single stimulus. Jackson (1977) proposed a more complex measure of response consistency that allows for the responses to multiple items to contribute to the consistency estimate at the expense of requiring each participant to complete many surveys. Scores on odd-numbered items associated with a particular scale

J. A. MacDonald (✉) · D. Trafimow
Department of Psychology, New Mexico State University,
PO Box 30001, MSC 3452, Las Cruces, NM 88003, USA
e-mail: jmacd@nmsu.edu

Published online: 23 February 2013

are summed, as are scores on the even-numbered items. This calculation is repeated across many scales, and the even- and odd-numbered sums are correlated to produce a consistency measure. This measure is sensitive to random response and specific factor errors but is unlikely to be affected by transient errors, unless transient errors manifest within the duration of a single survey.

In this brief report, we will discuss an effort to apply these concepts to measure the reliability (consistency) of a participant's responses in a more general decision-making task. In this context, participants typically perform a large number of trials across occasions (blocks of trials); reliability in this context addresses transient error of particular individuals, but it does not address their specific factor error. In fact, defining specific factor error in a decision-making context is not straightforward. Consider a traditional testing context where one has several items measuring attitude. Specific factor error is minimized to the extent that participants' responses on the different items are similar; that is, there is a high level of internal consistency. For example, participants with a positive attitude toward "eating chocolate" should give positive responses to "like–dislike" and "favorable–unfavorable" items, whereas participants with a negative attitude should give negative responses in response to these two types of items. In a typical detection or discrimination task, however, there are typically only two types of trials, with no expectation on the part of the researcher that responses to the two kinds of trials will be similar. For example, in a yes/no detection task, there is no obvious a priori reason for the researcher to believe that responses on target-present trials will be similar to responses on target-absent trials. If the participant is good at the task, the expectation might be for dissimilar responses: a "yes" response on target-present trials and a "no" response on target-absent trials. In this new context, internal consistency should concern not whether individuals respond similarly on all items measuring a construct but, rather, whether their responses to a particular trial type are similar. Our goal is to develop a measure of this similarity.

Derivation of the measure

Consider a decision-making task in which an observer is presented with one of two possible stimuli (A or B), and is required to identify it as either A or B. Yes/no detection, two-choice discrimination, and two-interval forced choice (2IFC) procedures all fit this description. Imagine that the experiment consists of two blocks of n trials. Each block comprises the same proportions of A and B stimuli, although the order of the stimuli varies across blocks.

A researcher interested in the consistency of the responses of a participant could consider the blocks as two administrations of the same test and correlate the responses across blocks to get a measure of test–retest reliability. Given that the stimulus order was allowed to vary across blocks, however, the researcher would first need to reorder the trials in the second block to match the stimulus order of the first. This would result in n pairs of trials, the response on each trial within a pair associated with either stimulus A or stimulus B. If the responses were recoded as numeric values (1 for A responses and 0 for B responses, for example), the correlation across blocks would provide a reasonable and intuitive measure of response consistency.

The object of this brief report is to derive an estimate of this quantity when only one block of trials is available. In other words, we are trying to predict test–retest reliability after a single administration of the test. One could interpret the measure as an indication of the consistency that would be observed if the participant completed the block of trials again while the percept distributions, stopping rule, and decision rule were held constant across blocks. We will start by identifying the formula for the correlation between the two blocks of responses. Since both sets of responses are dichotomous, each of the n trial pairs fits into one of four categories, as indicated in the following contingency table:

		Block 2 Trial	
		Response = A	Response = B
Block 1 Trial	Response = A	a	b
	Response = B	c	d

a, b, c, and d are the frequencies associated with each of the four possible outcomes, where $a + b + c + d = n$. The formula for the correlation between two dichotomous variables (also known as the phi coefficient) can be written in terms of these cell frequencies:

$$\rho = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{1}$$

The point of this derivation is to estimate ρ from quantities observed in block 1 under the assumption that the hit and false alarm rates (and therefore, the corresponding frequencies) are constant across blocks. We will also assume that the decision process is applied independently across blocks. We will denote the observed frequencies of hits, misses, false alarms, and correct rejections in block 1 as h, m, f, and r, respectively.

To begin with, we will derive the formula for the quantity d, which is the number of trials for which a B response was observed in both blocks. This combination of responses can happen only if a hit occurs in both blocks or if a false alarm occurs in both blocks. Using $\frac{h}{h+m}$ and $\frac{f}{f+r}$ as our estimates of the hit and false alarm rates, respectively, it follows that

$$d = h \cdot \frac{h}{h+m} + f \cdot \frac{f}{f+r} = \frac{h^2}{h+m} + \frac{f^2}{f+r}. \qquad (2)$$

c is a count of trials for which a hit occurred in the first block and a miss occurred in the second or a false alarm occurred in the first block and a correct rejection occurred in the second. Therefore,

$$c = h \cdot \frac{m}{h+m} + f \cdot \frac{r}{f+r} = \frac{hm}{h+m} + \frac{fr}{f+r}, \qquad (3)$$

and using the same logic,

$$b = m \cdot \frac{h}{h+m} + r \cdot \frac{f}{f+r} = \frac{hm}{h+m} + \frac{fr}{f+r} = c \qquad (4)$$

Finally,

$$a = \frac{m^2}{h+m} + \frac{r^2}{f+r}. \qquad (5)$$

All that remains is to insert these quantities into Eq. 1 and simplify terms. Doing so to the numerator of Eq. 1 yields

$$ad - bc = \frac{1}{(f+r)(h+m)} \cdot (hr - fm)^2. \qquad (6)$$

Moving on to the denominator of Eq. 1, given the result that *b* and *c* are equal, $(a + b) = (a + c)$ and $(b + d) = (c + d)$. Therefore,

$$\sqrt{(a+b)(c+d)(a+c)(b+d)} = (a+b)(c+d). \qquad (7)$$

$(a + b)$ is the number of A responses in block 1, and $(c + d)$ is the number of B responses in block 1. Writing these quantities in terms of *h*, *m*, *f*, and *r*, it follows that

$$(a+b)(c+d) = (m+r)(h+f). \qquad (8)$$

Inserting Eqs. 6 and 8 into Eq. 1, we get

$$\rho = \frac{(hr - fm)^2}{(f+r)(h+m)(m+r)(h+f)}. \qquad (9)$$

Equation 9 can also be expressed in terms of the hit rate, false alarm rate, the base rates, and the response rates:

$$\rho = \frac{P(S=A)P(S=B)}{P(R=A)P(R=B)}(HR - FAR)^2. \qquad (10)$$

In the above equation, $P\ (S = A)$ and $P\ (S = B)$ are the stimulus base rates, $P\ (R = A)$ and $P\ (R = B)$ are the observed proportions of A and B responses, and HR and FAR are the hit and false alarm rates, respectively. Please see the Appendix for the derivation of Eq. 10 from Eq. 9.

The relation between ρ and accuracy is straightforward to calculate. We define accuracy as the average of the hit and correct rejection rates:

$$Accuracy = \frac{HR + (1 - FAR)}{2}. \qquad (11)$$

Rearranging terms, we get

$$(HR - FAR) = 2(Accuracy - 0.5). \qquad (12)$$

Substituting into Eq. 10,

$$\rho = \frac{P(S=A)P(S=B)}{P(R=A)P(R=B)}(2(Accuracy - 0.5))^2. \qquad (14)$$

Therefore, if the stimulus and response probabilities are treated as constants,[1] consistency is proportional to the squared deviation of observed accuracy from 0.5:

$$\rho \propto (Accuracy - 0.5)^2. \qquad (15)$$

This result is in line with our expectation: Deviation from 50 % accuracy in either direction indicates that the responses are becoming more similar to one another.

A demonstration with real data

We will demonstrate the use and interpretation of Eq. 9 using data from Experiment 1b from Trafimow, MacDonald, and Rice (2012). The experiment consisted of a 2IFC auditory detection task made up of two blocks of 52 trials each. For the purposes of this demonstration, we will treat the 2IFC task as an *AB* discrimination task: Stimulus *A* is the concatenation of a signal interval with a noise interval, and stimulus *B* is the concatenation of a noise interval with a signal interval. In this context, a "hit" occurs when stimulus *B* is presented (the signal occurs in the second interval) and the participant responds *B,* and a correct rejection occurs when stimulus A is presented (the signal occurs in the first interval) and the participant responds *A*. Response consistency estimates were obtained for each participant using Eq. 9 and data from the first block of trials. The actual response consistency was calculated for each participant by matching trials by stimulus type across blocks and correlating the responses. Although 31 participants completed the experiment, 2 were excluded from analysis due to response behavior that resulted in a value of zero for the denominator of Eq. 9.

The results are illustrated in Fig. 1. The dotted line indicates a perfect correspondence between estimated and observed consistency. The consistency estimates were quite accurate: They accounted for 90.8 % of the variance in the observed consistency scores. As can be seen in the figure, there was a tendency to underestimate consistency. This presumably occurred because the decision process is likely to be unstable when the participant is learning how to complete the task, which would lead to a reduced consistency estimate. The decision process is more likely to be stable during the second block, which explains why the

[1] This treatment is not strictly appropriate, since the response probabilities are related to the hit and false alarm rates and are, therefore, likely to vary with the hit and false alarm rates.

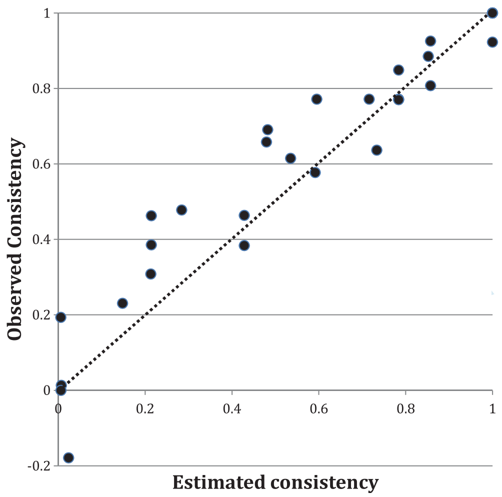

Fig. 1 A comparison of predicted and actual response consistency across blocks. The dotted line indicates a perfect correspondence between predicted and observed consistency. Points above the line represent underestimates, and points below represent overestimates

observed consistency is higher than the estimate produced from the first block of data.

Discussion

The proposed measure provides an estimate of the correlation between the responses associated with two iterations of the same decision-making experiment. The measure identifies the consistency that would be observed if its assumptions were met. First, we assume that the decision process that maps stimuli to responses is independent across blocks. In other words, we assume that if the stimulus type is held constant, the response probabilities are identical across blocks. The validity of this assumption is easily evaluated using standard tests of independence. This assumption is likely to be invalid during the early phases of an experiment where the decision strategy of the observer is still under development. Any change in the way in which percepts are matched to responses (i.e., any difference in the decision rule across blocks) will result in a violation of this assumption. For this reason, we advocate for the measure being interpreted as an indication of response consistency given a particular decision strategy, rather than an unqualified estimate of future behavior.

It follows from this assumption that the hit and false alarm rates will be constant across blocks. For this reason, the hit and false alarm rates observed during the first trial block are used as point estimates of the hit and false alarm rates in the second (upcoming) block of trials. Given that these quantities are sample statistics, they are very unlikely to be equal across blocks even if the decision process that produced them remains constant. If the hit and false alarm rates in the second block were to deviate significantly from those observed in the first, the accuracy of the consistency estimate would suffer. However, it is clear from the demonstration illustrated in Fig. 1 that this occurred only rarely, at least for the data set under consideration: The mean absolute error in predicted consistency was 0.083 across 29 participants, and the maximum error in prediction was only 0.187. In any case, such a deviation across blocks would indicate that the decision process had not stabilized.

As Cronbach (1947) explained, no measure of reliability is perfect, and each one offers different benefits. So it is useful to contrast them against each other. Most measures of reliability, such as Cronbach's alpha and the Spearman–Brown formula, each test the internal consistency of items across groups of participants. But if a researcher wishes to obtain the reliability with which individuals respond to two sets of stimuli, none of these methods are applicable in a straightforward way. In an attempt to fill this gap, a researcher could obtain reliability coefficients for individuals by having them respond to at least two blocks of matched trials. This might be considered a within-participants version of test–retest reliability. Although this measure often works well (e.g., Trafimow & Rice, 2009), it cannot always be used. Sometimes the experimental design requires that there is only one block of trials, or the measure may be applied after the fact to a preexisting data set. Even if it is possible to have two blocks of trials, perhaps there is no way to match the trials on the first block with the trials on the second block. Without such matching, there is no way to compute reliability across blocks of trials. Our proposed consistency measure addresses this problem by providing an estimate, from a single block of trials, of the result that likely would occur with two blocks of trials. Although the accuracy of the estimate depends on the assumptions we discussed earlier, the analyses illustrated in Fig. 1 suggest that there is reason to be hopeful that the assumptions are not strongly violated in at least some experimental contexts.

It is possible to argue that unless our proposed consistency measure actually is tested against a two-block measure, as is illustrated in Fig. 1, it is difficult to know whether to trust it. And if it is necessary to use two blocks of trials anyhow, why not simply compute a within-participants version of test–retest reliability and not bother with the proposed consistency measure? We suggest multiple answers. In the first place, the proposed consistency measure can be used on each block of trials to provide a test of changes in consistency across blocks of trials, possibly due to learning, exhaustion, or other factors. Theory-based predictions about changes in consistency could therefore be tested using this measure. Second, in many areas in perceptual and cognitive psychology, researchers tend to produce large numbers of articles using very similar research paradigms. Once preliminary research has been accomplished and the proposed consistency measure has been checked against a two-block measure and validated for a particular research paradigm, it is inefficient for each researcher in the future to continue to use a two-block paradigm where a one-block paradigm, in concert with our proposed consistency measure, is sufficient. Therefore, given appropriate early testing, our proposed consistency measure can reduce the difficulty of performing research and, thereby, increase the ratio of knowledge gained to effort put forth.

We see an additional use of the ρ measure that distinguishes it from test–retest reliability across blocks of trials. Suppose that a researcher presents participants with two types of trials and is interested in comparing the internal consistency of each participant's responses to each type of trial. Prior to the development of the ρ measure, there was no index that provides this capability. But ρ can be applied separately to each of the two trial types. For example, suppose visual stimuli were presented with different contrasts, and there was a theoretical reason to predict that people would respond with greater internal consistency to trials with one contrast than to trials with the other. Equation 9 could be used separately, for the trials within each contrast type, to obtain internal consistency coefficients for each. A comparison of the two coefficients with predictions would provide a test of the theory.

In summary, we see ρ as useful for obtaining a reliability index for individuals when it is impossible or impractical to obtain data across blocks of trials. In addition, ρ enables researchers to compare the consistency of a person's responses to different types of trials. If the researcher arranges matters so that the different types of trials map on to predictions from a theory, ρ can be used to test the predictions. In short, we hope that the ρ measure will be a useful addition to the literature concerning the assessment of the reliability of particular individuals in a decision-making context.

Appendix

In this Appendix, we will derive Eq. 10 from Eq. 9 in the main text. Starting with Eq. 9:

$$\rho = \frac{(hr - fm)^2}{(f + r)(h + m)(m + r)(h + f)}.$$

Add and subtract hf within the squared term in the numerator:

$$= \frac{(hf + hr - hf - fm)^2}{(f + r)(h + m)(m + r)(h + f)}.$$

Factor out a couple of common terms in the numerator:

$$= \frac{[h(f + r) - f(h + m))]^2}{(f + r)(h + m)(m + r)(h + f)}.$$

Multiply the numerator and denominator by $\frac{1}{(f+r)^2(h+m)^2}$:

$$= \frac{\frac{1}{(f+r)^2(h+m)^2}}{\frac{1}{(f+r)^2(h+m)^2}} \cdot \frac{[h(f+r) - f(h+m))]^2}{(f+r)(h+m)(m+r)(h+f)}$$
$$= \frac{\frac{[h(f+r) - f(h+m))]^2}{(f+r)^2(h+m)^2}}{\frac{(m+r)(h+f)}{(f+r)(h+m)}}$$
$$= \frac{(f+r)(h+m)}{(m+r)(h+f)} \cdot \left[\frac{h(f+r) - f(h+m))}{(f+r)(h+m)}\right]^2.$$

Multiply the numerator and denominator by $\frac{1}{n^2}$, where n is the number of trials in the block:

$$= \frac{\frac{1}{n^2}}{\frac{1}{n^2}} \cdot \frac{(f + r)(h + m)}{(m + r)(h + f)} \cdot \left[\frac{h}{h + m} - \frac{f}{f + r}\right]^2.$$

$\frac{h}{h+m}$ and $\frac{f}{f+r}$ are the hit and false alarm rates, respectively:

$$= \frac{\frac{f+r}{n} \cdot \frac{h+m}{n}}{\frac{m+r}{n} \cdot \frac{h+f}{n}} \cdot (HR - FAR)^2$$
$$= \frac{P(S = A) \cdot P(S = B)}{P(R = A) \cdot P(R = B)} \cdot (HR - FAR)^2$$

References

Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika, 12,* 1–16.

Jackson, D. N. (1977). *Jackson Vocational Interest Survey: Manual*. Port Huron, MI: Sigma Assessment Systems.

Nesselroade, J. R., & Ram, N. (2004). Studying intraindividual variability: What we have learned that will help us understand lives in context. *Research in Human Development, 1,* 9–29.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8,* 206–224.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15,* 72–101.

Trafimow, D., MacDonald, J. A., & Rice, S. (2012). Using PPT to account for randomness in perception. *Attention, Perception, & Psychophysics, 74,* 1355–1365.

Trafimow, D., & Rice, S. (2009). Potential Performance Theory (PPT): Describing a methodology for analyzing task performance. *Behavior Research Methods, 41,* 359–371.