Biometrics WILEY

Generalized reliability based on distances

Meng Xu[1] | **Philip T. Reiss**[1] | **Ivor Cribben**[2]

[1]Department of Statistics, University of Haifa, Haifa, Israel

[2]Department of Accounting, Operations, and Information Systems, University of Alberta School of Business, Edmonton, Canada

Correspondence
Philip T. Reiss, Department of Statistics, University of Haifa, Haifa 31905, Israel.
Email: reiss@stat.haifa.ac.il

Abstract

The intraclass correlation coefficient (ICC) is a classical index of measurement reliability. With the advent of new and complex types of data for which the ICC is not defined, there is a need for new ways to assess reliability. To meet this need, we propose a new distance-based ICC (dbICC), defined in terms of arbitrary distances among observations. We introduce a bias correction to improve the coverage of bootstrap confidence intervals for the dbICC, and demonstrate its efficacy via simulation. We illustrate the proposed method by analyzing the test-retest reliability of brain connectivity matrices derived from a set of repeated functional magnetic resonance imaging scans. The Spearman-Brown formula, which shows how more intensive measurement increases reliability, is extended to encompass the dbICC.

KEYWORDS

functional connectivity, intraclass correlation coefficient, Spearman-Brown formula, test-retest reliability

1 | INTRODUCTION

With the increasing availability of new and complex forms of data, there is a corresponding need for new ways to assess measurement reliability. This article aims to help meet this need by reformulating the intraclass correlation coefficient (ICC), a standard index of reliability, in terms of distances between observations.

We begin by defining the ICC as developed in classical test theory (Lord and Novick, 1968; Fleiss, 1986; Mair, 2018), which views a measured scalar quantity X as the sum of an underlying true score T and an error term E. Suppose we have a sample of I individuals with true real-valued scores $T_1, \ldots, T_I$ drawn from a population with variance σ_T^2; and that for each i, the ith individual is measured J_i times, yielding observations

$$X_{ij} = T_i + \varepsilon_{ij}, \tag{1}$$

$j = 1, \ldots, J_i$, where the ε_{ij}'s are drawn from a distribution with mean 0 and variance σ_ε^2, independently of each other and of the T_i's. Then for distinct $j_1, j_2 \in \{1, \ldots, J_i\}$, the correla-

tion between the j_1th and j_2th observations for individual i is easily shown to be

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\varepsilon^2}. \tag{2}$$

This quantity is the classical ICC.

Reliability measures for more complex settings include replacing model (1) with the generalizability theory model of Cranford *et al.* (2006), as well as generalizations of (2) to multivariate data (Alonso *et al.*, 2010), including high-dimensional data (Shou *et al.*, 2013). All of these extensions assume a model that is more complex than (1), but still of an additive (signal plus noise) form. However, for complex objects that are measured or estimated in modern biomedical research, such as motion patterns or brain networks, such an additive representation is typically inapplicable. There is thus a need for a new reliability index appropriate for general data objects.

Our work was motivated by the study of functional connectivity in the human brain by means of resting-state functional
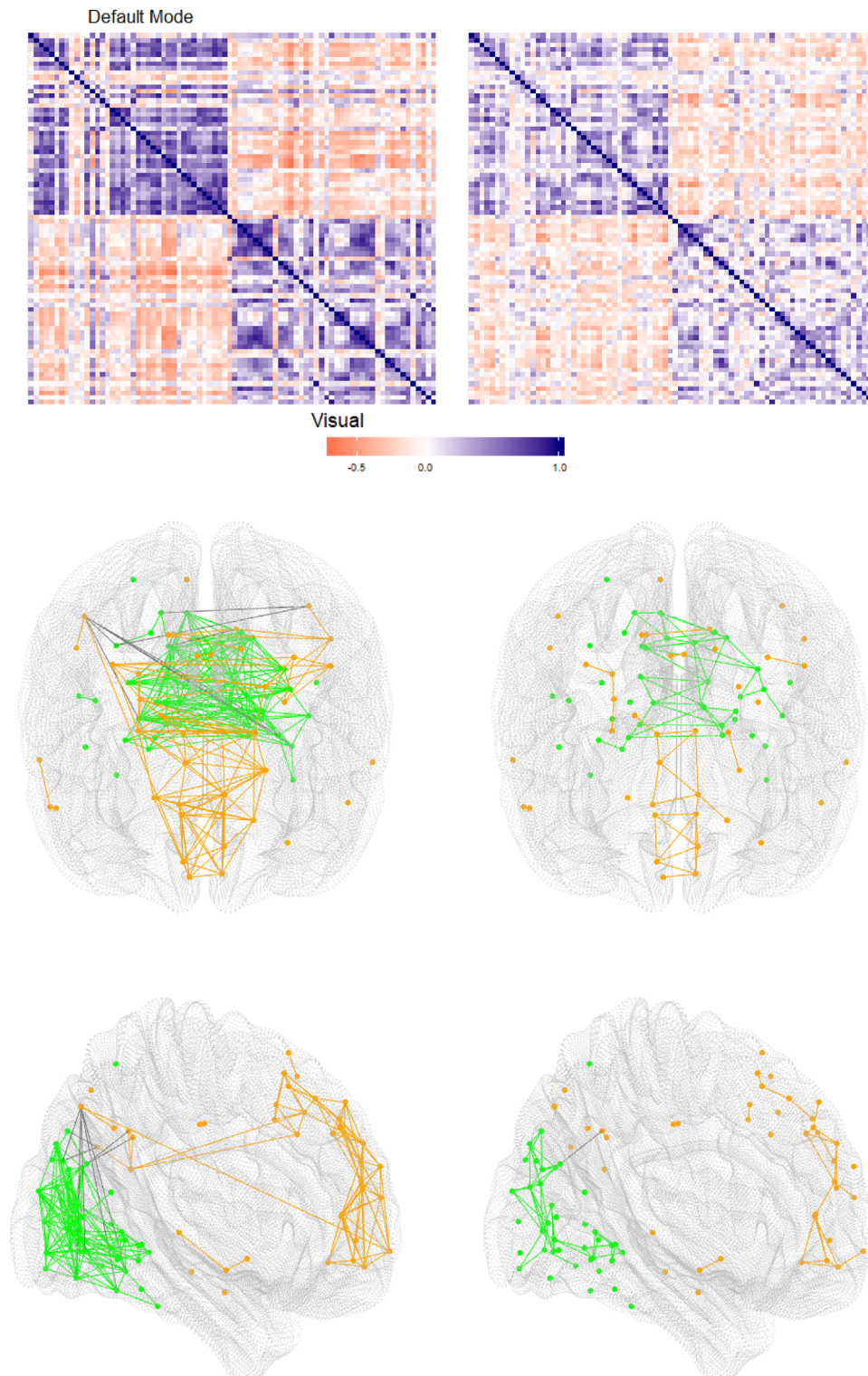
FIGURE 1 *Top:* Matrices $\boldsymbol{R}$ of correlations among 80 ROIs comprising the default mode network and visual network in our fMRI data set. The left and right matrices, respectively, attain the highest and lowest connectivity scores $-\log|\boldsymbol{R}|$ observed in our data set. *Middle:* Brain maps (axial view) corresponding to the same two correlation matrices, and displaying pairs of regions with absolute correlation above 0.6. Orange nodes and links refer to the default mode network; green nodes and links refer to the visual network; links between the two networks are shown in black. *Bottom:* Same brain maps, sagittal view. The fMRI data are presented in Section 4, and the connectivity score $-\log|\boldsymbol{R}|$ is discussed briefly in Web Appendix A

magnetic resonance imaging (fMRI). Briefly, fMRI produces a time series of brain activity, known as the blood oxygen-level-dependent (BOLD) signal, at each of a set of regions of interest (ROIs). *Resting-state* fMRI means that the participants in the study were not performing any particular task or viewing a stimulus during the brain scan. Functional connectivity refers to association among activity levels in different parts of the brain, and can be measured in many ways (Yan *et al.*, 2013). One of the most common functional connectivity measures is a simple Pearson correlation matrix of regional BOLD signals. Figure 1 displays two such correlation matrices, along with associated brain graphs, for a set of 80 ROIs to be discussed in Section 4. These particular examples were chosen to illustrate high and low connectivity, according to a metric described in Web Appendix A.

In order to be confident that such correlation matrices, and the scientific conclusions derived from them, are trustworthy and reproducible, it is necessary first to be able to assess their reliability (Noble *et al.*, 2019). Our proposed methodology offers a means to that end.

Our basic proposal, a reformulation of the ICC based on distances between observations, is outlined in Section 2, and estimation of the resulting reliability index is discussed in Section 3. An application to an fMRI data set is presented in Section 4. In Sections 5-7, we extend the Spearman-Brown (SB) formula, a fundamental result in reliability theory, to our distance-based ICC, and revisit our fMRI data set in light of this extension. A concluding discussion appears in Section 8.

2 | DISTANCE-BASED RELIABILITY MEASUREMENT

A novel reliability index applicable to general data objects can be defined by rederiving the ICC (2) in terms of squared distances among observations. Let $\mathrm{MSD}_b = E_{i_1 \neq i_2}[(X_{i_1 j_1} - X_{i_2 j_2})^2]$ and $\mathrm{MSD}_w = E_{j_1 \neq j_2}[(X_{i j_1} - X_{i j_2})^2]$ be the mean squared differences for measurements between and within individuals, respectively. Then $\mathrm{MSD}_b = 2\sigma_T^2 + 2\sigma_\epsilon^2$ and $\mathrm{MSD}_w = 2\sigma_\epsilon^2$, and thus, the ICC (2) can be reexpressed as

$$\rho = 1 - \frac{\mathrm{MSD}_w}{\mathrm{MSD}_b}. \tag{3}$$

The advantage of expression (3) is that, unlike (2), it extends straightforwardly to general data objects (curves, networks, etc), as long as a distance or dissimilarity $d(\cdot, \cdot)$ between such objects is defined. One simply redefines MSD_b and MSD_w in (3) in a more general sense, as the between- and within-individual mean squared *distances*

$$\mathrm{MSD}_b = E_{i_1 \neq i_2}\left[d\left(X_{i_1 j_1}, X_{i_2 j_2} \right)^2 \right] \quad \text{and}$$

$$\mathrm{MSD}_w = E_{j_1 \neq j_2}\left[d\left(X_{i j_1}, X_{i j_2} \right)^2 \right]. \tag{4}$$

Henceforth, we shall refer to (3), with $\mathrm{MSD}_b, \mathrm{MSD}_w$ given by (4), as the *distance-based intraclass correlation coefficient*, or dbICC.

We note that the same general strategy, of rederiving variance-based formulas in terms of sums of squared distances, has previously been used to formulate distance-based hypothesis tests (McArdle and Anderson, 2001; Mielke and Berry, 2007; Reiss *et al.*, 2010).

A simple example of extending (1) beyond the scalar real-valued case is to let T_i, ϵ_{ij} be mutually independent random *vectors*, with covariance matrices $\Sigma_T, \Sigma_\epsilon$, respectively, and let d be the Euclidean distance. Then (3) reduces straightforwardly to

$$\rho = 1 - \frac{\mathrm{tr}(\Sigma_\epsilon)}{\mathrm{tr}(\Sigma_T + \Sigma_\epsilon)} = \frac{\mathrm{tr}(\Sigma_T)}{\mathrm{tr}(\Sigma_T + \Sigma_\epsilon)}, \tag{5}$$

the multivariate reliability measure referred to as R_T (Alonso *et al.*, 2010), and as I2C2 (Shou *et al.*, 2013) for images viewed as vectors. Thus, the dbICC is an extension of these measures to more general distances and data types.

3 | ESTIMATING THE dbICC

3.1 | Point estimation

Like the classical ICC (2), the proposed dbICC (3) can be estimated in practice by plugging in consistent estimates of the population quantities (4), as follows:

$$\hat{\rho} = 1 - \frac{\widehat{\mathrm{MSD}}_w}{\widehat{\mathrm{MSD}}_b}, \tag{6}$$

where

$$\widehat{\mathrm{MSD}}_b = \frac{\sum_{1 \leq i_1 < i_2 \leq I} \sum_{j_1=1}^{J_{i_1}} \sum_{j_2=1}^{J_{i_2}} d(X_{i_1 j_1}, X_{i_2 j_2})^2}{\sum_{1 \leq i_1 < i_2 \leq I} J_{i_1} J_{i_2}}, \tag{7}$$

$$\widehat{\mathrm{MSD}}_w = \frac{\sum_{i=1}^{I} \sum_{1 \leq j_1 < j_2 \leq J_i} d(X_{i j_1}, X_{i j_2})^2}{\sum_{i=1}^{I} \binom{J_i}{2}}. \tag{8}$$

Figure 2 illustrates this schematically for a distance matrix with rows and columns grouped by individuals: one estimates $\mathrm{MSD}_b, \mathrm{MSD}_w$ by averaging the between- and within-individual distances (B and W), respectively.

3.2 | Bootstrap confidence intervals

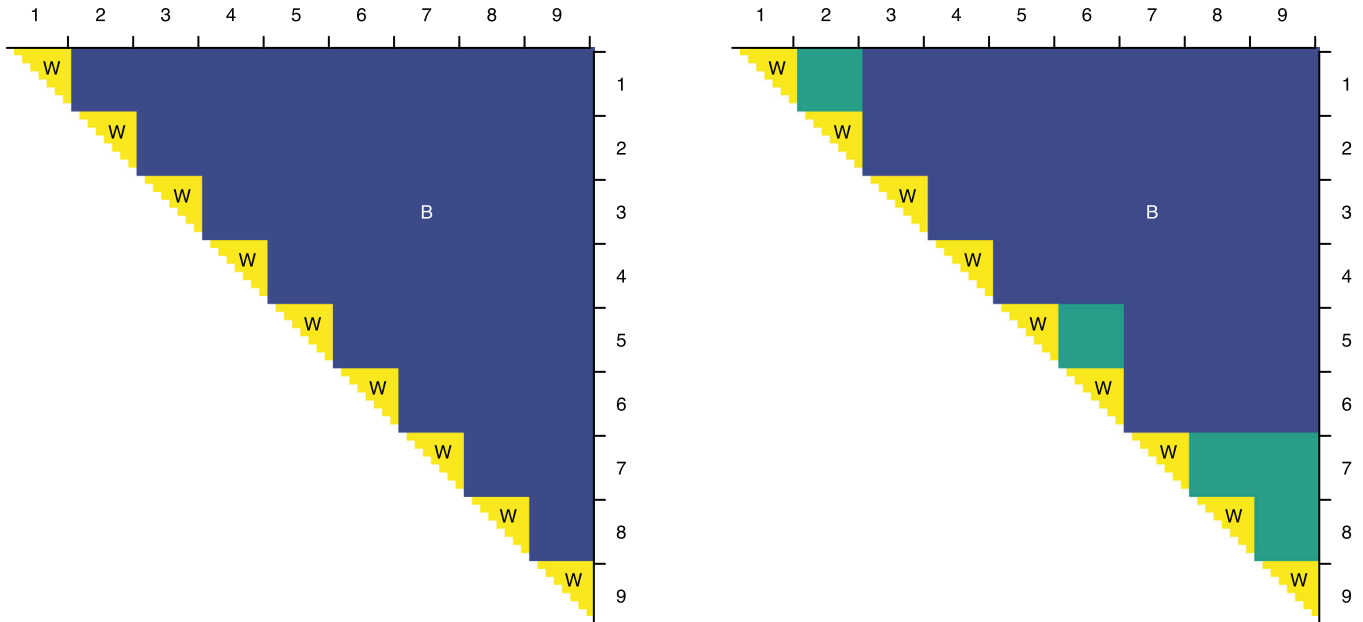The dbICC is intended for distance functions whose distribution may not be known. It is thus natural to turn to

FIGURE 2 *Left:* Schematic diagram of a matrix of distances among repeated observations of nine individuals, with rows and columns grouped by individual. Distances in the half-squares along the diagonal are within-individual (W), while the rest are between-individual (B). *Right:* A similar diagram, but for a bootstrap sample with repeated observations. Distances shown in green are nominally between-individual, but in reality, they are within-individual

nonparametric bootstrapping as a distribution-free approach to interval estimation for the dbICC. For $r = 1, \ldots, B$ with suitably large B, let $\pi_1^r, \ldots, \pi_I^r$ be a sample with replacement from $\{1, \ldots, I\}$; then the rth bootstrap sample consists of $X_{ij}^r \equiv X_{\pi_i^r j}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J_{\pi_i^r}$. The resulting ICC estimate is

$$\hat{\rho}^r = 1 - \frac{\widehat{\text{MSD}}_w^r}{\widehat{\text{MSD}}_b^r}, \tag{9}$$

where $\widehat{\text{MSD}}_w^r, \widehat{\text{MSD}}_b^r$ are bootstrap analogues of (7) and (8):

$$\widehat{\text{MSD}}_b^r = \frac{\sum_{1 \leq i_1 < i_2 \leq I} \sum_{j_1=1}^{J_{\pi_{i_1}^r}} \sum_{j_2=1}^{J_{\pi_{i_2}^r}} d(X_{i_1 j_1}^r, X_{i_2 j_2}^r)^2}{\sum_{1 \leq i_1 < i_2 \leq I} J_{\pi_{i_1}^r} J_{\pi_{i_2}^r}}, \tag{10}$$

$$\widehat{\text{MSD}}_w^r = \frac{\sum_{i=1}^{I} \sum_{1 \leq j_1 < j_2 \leq J_{\pi_i^r}} d(X_{ij_1}^r, X_{ij_2}^r)^2}{\sum_{i=1}^{I} \binom{J_{\pi_i^r}}{2}}.$$

The interval from the $\alpha/2$ to the $1 - \alpha/2$ quantile of the $\hat{\rho}^r$'s can then be used as a $100(1 - \alpha)\%$ confidence interval.

These bootstrap estimates $\hat{\rho}^r$, however, suffer from negative bias (over and above the well-known negative bias of the classical ICC; Atenafu *et al.*, 2012). Returning to the example in Figure 2, consider a bootstrap sample in which individuals 1 and 2 are duplicates, as are individuals 5 and 6 and individuals 7-9. Then the blocks shown in the right subfigure in green

nominally refer to between-individual differences, but, in fact, represent within-individual differences. Assuming $\text{MSD}_w < \text{MSD}_b$, counting these entries as between-individual will tend to result in underestimation of MSD_b and hence in negative bias in (9). The diagonal entries of these blocks are zero, thereby compounding the bias. To remove this bias, we can simply exclude such blocks from the summations in (10); formally, we replace each occurrence of $\sum_{1 \leq i_1 < i_2 \leq I}$ with $\sum_{1 \leq i_1 < i_2 \leq I, \pi_{i_1}^r \neq \pi_{i_2}^r}$.

3.3 | A simulation study

Using multivariate data with Euclidean distance (the example from the end of Section 2), we conducted a simulation study to assess the accuracy of our point and interval estimates of the dbICC. Values $X_{ij} \in \mathbb{R}^2$ were drawn from (1) where $T_i \sim N_2(0, I_2)$ and $\varepsilon_{ij} \sim N_2(0, c I_2)$ with $c = 4, 1, 0.25$. By (5), the (population) dbICC is then $\rho = \frac{1}{c+1}$, which equals 0.2, 0.5, and 0.8 for the above three values of c. The number of subjects I was set to 10, 40, and 70, and the number of measurements per subject J_i fixed at 4. We took 500 replicates with each combination of the above values of ρ and I. Boxplots of the dbICC estimates are displayed in Figure 3. The classical negative bias of ICC estimates (Atenafu *et al.*, 2012) is noticeable for $I = 10$ when $\rho = 0.2, 0.5$, but not for the other settings.

Next, we considered bootstrap confidence intervals, with $B = 1200$, without and with the bias correction of the
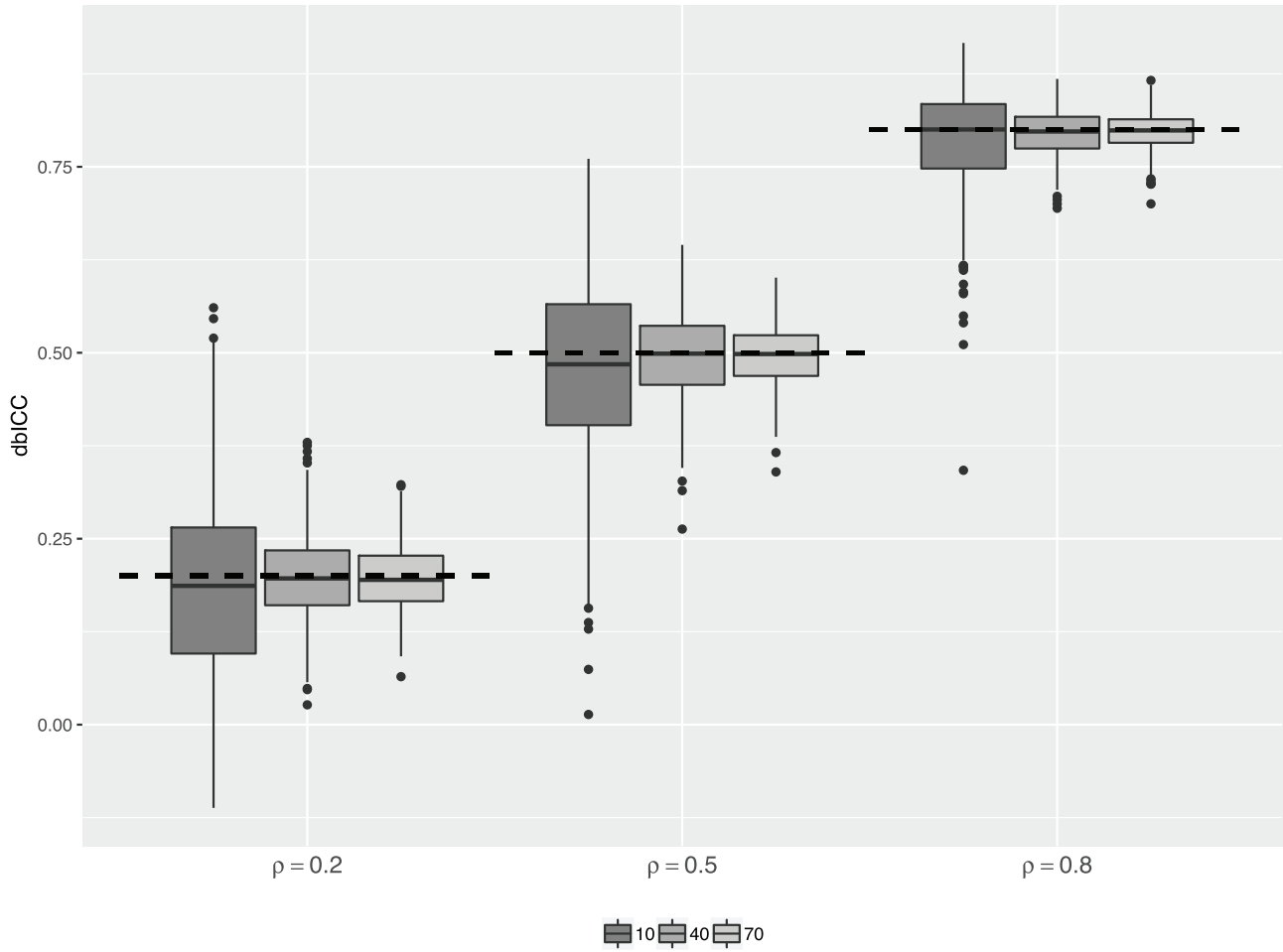
FIGURE 3 Boxplots of point estimates of dbICC, for true values $\rho = 0.2, 0.5, 0.8$ (indicated by dashed lines) and for $I = 10, 40, 70$
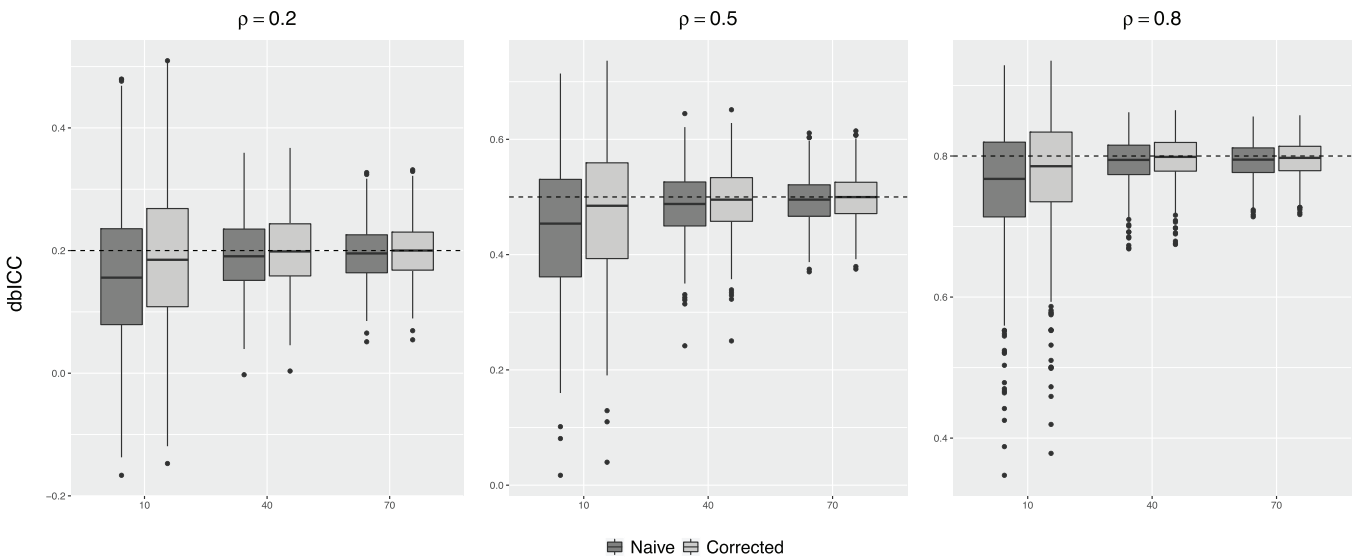

FIGURE 4 Boxplots of median bootstrap estimate of dbICC, for true values $\rho = 0.2, 0.5, 0.8$ (indicated by dashed lines) and for $I = 10, 40, 70$

TABLE 1 Percent coverage of bootstrap 95% confidence intervals, naïve (N) and corrected (C)

	$I = 10$		$I = 40$		$I = 70$	
	N	C	N	C	N	C
$\rho = 0.2$	86.0	90.8	91.6	93.2	92.2	92.6
$\rho = 0.5$	84.8	90.6	91.4	92.0	94.0	94.6
$\rho = 0.8$	85.2	89.6	90.6	92.6	92.8	94.2

TABLE 2 Point estimates and 95% bootstrap CIs for dbICC, based on three sets of ROIs and three distance measures

	ℓ_2	ℓ_1	$\sqrt{1-r}$
All 333 ROIs	0.378	0.382	0.382
	(0.329,0.424)	(0.335,0.426)	(0.338,0.426)
Default mode network	0.488	0.493	0.487
	(0.403,0.562)	(0.404,0.570)	(0.414,0.555)
Visual network	0.434	0.435	0.451
	(0.362,0.508)	(0.354,0.515)	(0.401,0.500)

previous subsection. We performed 500 replicates for each combination of the same ρ and I values as above, again with J_i fixed at 4. Boxplots of the median of the 1200 bootstrap estimates within each replicate are presented in Figure 4. For $I = 10$ and to some extent for $I = 40$, the correction yields a marked reduction in the observed negative bias. Accordingly, the coverage of 95% confidence intervals is improved by the correction, as can be seen in Table 1. As noted above, however, a small-sample negative bias (unrelated to bootstrapping) occurs for point estimates of dbICC as for the classical ICC, and hence the coverage remains quite poor for $I = 10$.

4 | FUNCTIONAL CONNECTIVITY IN THE HUMAN BRAIN

As noted in the introduction, the dbICC was originally conceived as a way to evaluate the reliability of functional connectivity measures. To demonstrate how dbICC can be so applied, here we reexamine part of a data set presented by Shehzad *et al.* (2009) in an early study of the test-retest reliability of resting-state functional connectivity. These authors, followed by others (eg, Somandepalli *et al.*, 2015; Choe *et al.*, 2017), focused on ordinary ICC at each of a set of brain locations or connections. The dbICC, by contrast, offers an overall index of reliability for fMRI-based correlation matrices, viewed as *gestalt* measures of functional connectivity.

The data include BOLD time series of length 197, within each of 333 ROIs derived by Gordon *et al.* (2016), for $I = 25$ individuals, with $J = 2$ such fMRI scans per individual; further details are provided in the Appendix. We then computed the distance between each pair of matrices R_1, R_2 among the $25 \times 2 = 50$ correlation matrices thus derived, using each of three distance measures:

(i) the ℓ_2 distance (square root of sum of squared differences) between vec(R_1) and vec(R_2);

(ii) the ℓ_1 distance (sum of absolute differences) between vec(R_1) and vec(R_2); and

(iii) $\sqrt{1-r}$, where r is the correlation between the lower triangular elements of R_1 and those of R_2 (correlation of correlations); the rationale for this distance is explained in Web Appendix B.

We stress that (i) and (ii) are not the distances induced by the matrix 2- and 1-norms, since here we are interested in entry-wise differences as opposed to treating the matrices as operators. Distance (i) is, rather, the distance induced by the Frobenius norm, which, in turn, is induced by an inner product; consequently, this distance fits with the generalized true score model presented below in Section 5.2. Since the matrices are treated here as vectors, dbICC based on distance (i) is equivalent to the I2C2 estimator of Shou *et al.* (2013) cited at the end of Section 2, although these authors focused on MRI-based images as opposed to regional connectivity matrices.

The dbICC estimates (6) based on distances (i)-(iii), along with 95% bootstrap CIs, are given in the first row of Table 2. While fairly consistent with the results of Shou *et al.* (2013), these reliabilities are very low by classical standards.

We also examined two subsets of the 333 ROIs: 41 ROIs constituting the *default mode network* of the brain (DMN; Raichle *et al.*, 2001), and 39 ROIs making up the brain's visual network. Correlations among the ROIs within each of these networks tend to be high, as illustrated in Figure 1. Hence, it comes as no surprise that dbICC values within each of these two networks, presented in the second and third rows of Table 2, are markedly higher than for the complete set of ROIs. For each set of ROIs, the dbICC values are quite consistent across the three distances.

A likely explanation for the relatively low dbICCs for the complete set of 333 ROIs is that many pairs of regions are essentially uncorrelated, and thus, their correlation estimates largely reflect noise. This suggests that it might be possible to boost dbICC by thresholding small correlations. Figure 5 shows the effect on dbICC of soft-thresholding. Somewhat contrary to our expectation, soft-thresholding generally increased dbICC only slightly at best, and often decreased it.

5 | GENERALIZING THE SPEARMAN-BROWN FORMULA

Is there a way to improve upon the low reliabilities found for the functional connectivity data? A general approach to boosting reliability, suggested by classical psychometrics, is to take more measurements: for example, to average over replicates
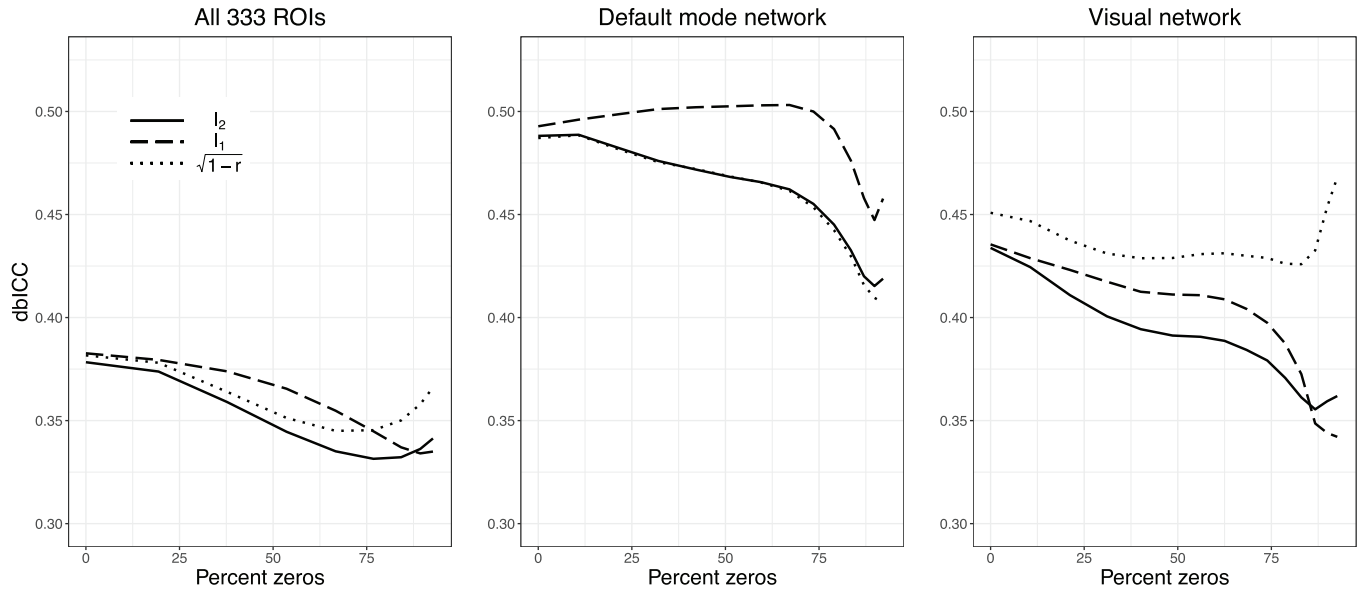
FIGURE 5 Estimated dbICC, for the same distances and sets of ROIs as in Table 2, but with soft-thresholding of the correlation values. The horizontal axis denotes the average percentage of the correlations that are shrunk to zero, as the threshold increases

of a measure, or to increase the number of questions on a test. A well-known relation between the number of measurements and the reliability appeared in Spearman (1910) and, in a more familiar form, in Brown (1910). In this section, we extend this relation to the distance-based ICC, and in Section 6, we reexamine the fMRI data results in light of our generalization of the Spearman-Brown (SB) formula.

5.1 | Measurement intensity and its effect on reliability

The SB formula states that averaging each score over m replicates transforms the classical ICC from ρ to $m\rho/[1 + (m - 1)\rho]$. If we let ρ_1, ρ_m, respectively, denote the raw ICC and the ICC based on m replicates, the formula can be written as $\rho_m = m\rho_1/[1 + (m - 1)\rho_1]$, which with some rearrangement becomes

$$\frac{\rho_m}{1 - \rho_m} = m\frac{\rho_1}{1 - \rho_1},$$

or alternatively

$$\frac{\rho_m}{1 - \rho_m} \propto m. \tag{11}$$

Lord and Novick (1968) refer to $\rho/(1 - \rho)$ as the signal-to-noise ratio (SNR), and accordingly, (11) may be paraphrased as: the SNR is proportional to the number of measurements whose average is taken.

Averaging over m real-valued measurements can be viewed as just one example of a broader notion of increasing *mea-surement intensity* and thereby boosting reliability. Other instances of measurement intensity m include:

(E1) An estimated covariance or correlation matrix based on a sample of m multivariate observations. For functional connectivity matrices as considered above in Section 4, m would be the number of time points recorded by fMRI.

(E2) A curve estimate obtained by penalized spline smoothing with m observations.

Our goal in the next subsection is to derive a distance-based SB relation, ie, an analogue of (11) in which m denotes measurement intensity and ρ_m is the resulting dbICC. To do this, we need a more general formulation of the true score model (1).

5.2 | A true score model for general Hilbert spaces

The classical setting of real-valued measures, as well as examples (E1) and (E2), can all be viewed as instances of a general setup in which the observations are of the form (1), but the T_i's are a random sample of true scores in a Hilbert space $\mathcal{H}$, while the ε_{ij}'s are random measurement errors in $\mathcal{H}$. We define distance in $\mathcal{H}$ by $d(h_1, h_2) = \|h_1 - h_2\|$, where $\|\cdot\|$ is the norm induced by the inner product on $\mathcal{H}$. Define

$$\Delta_T = E\left(\|T_{i_1} - T_{i_2}\|^2\right) \tag{12}$$

and

$$\Delta_\varepsilon(m) = E_m\left(\|\varepsilon_{i_1 j_1} - \varepsilon_{i_2 j_2}\|^2\right), \qquad (13)$$

for $i_1, i_2 \in \{1, \dots, I\}$ and $j_k \in \{1, \dots, J_k\}$ for $k = 1, 2$, where E_m denotes expectation for measurement intensity equal to m. Note that the measurement intensity affects only the expected distance between errors ε_{ij}, but not that between scores T_i. We make two assumptions, of which the first is implicit in (13):

(a1) The expectation in (13) is the same for $i_1 = i_2$ versus for $i_1 \neq i_2$.
(a2) For all i_1, i_2, j_1, j_2,

$$E(\langle T_{i_1} - T_{i_2}, \varepsilon_{i_1 j_1} - \varepsilon_{i_2 j_2}\rangle) = 0. \qquad (14)$$

Then

$$\begin{aligned} \rho_m &= 1 - \frac{E(\|X_{ij_1} - X_{ij_2}\|^2)}{E(\|X_{i_1 j_1} - X_{i_2 j_2}\|^2)} \\ &= 1 - \frac{E(\|\varepsilon_{ij_1} - \varepsilon_{ij_2}\|^2)}{E(\|T_{i_1} + \varepsilon_{i_1 j_1} - T_{i_2} - \varepsilon_{i_2 j_2}\|^2)} \\ &= 1 - \frac{\Delta_\varepsilon(m)}{\Delta_T + \Delta_\varepsilon(m)} \text{ [by (12),(13),(14)]} \\ &= \frac{\Delta_T}{\Delta_T + \Delta_\varepsilon(m)}, \end{aligned}$$

and therefore

$$\frac{\rho_m}{1 - \rho_m} = \frac{\Delta_T}{\Delta_\varepsilon(m)} \propto \frac{1}{\Delta_\varepsilon(m)}. \qquad (15)$$

In the classical case where X_{ij} is the mean of m measurements, ε_{ij} is the mean of m independent errors with mean 0 and common variance, so that

$$\Delta_\varepsilon(m) = E(\|\varepsilon_{i_1 j_1} - \varepsilon_{i_2 j_2}\|^2) \propto 1/m;$$

plugging this into (15) leads directly to the rearranged SB formula (11). In other cases, such as (E2), $\Delta_\varepsilon(m) \not\propto 1/m$, and hence, the generalized SB formula (15) does not reduce to (11).

6 | APPLYING THE GENERALIZED SB FORMULA TO THE fMRI DATA

Our goal in this section is to study the implications of the generalized SB formula (15) for correlation matrices such as those used in Section 4 as measures of functional connectivity. In Section 6.1 we show that, in the simpler setting of covariance matrix estimation, the relationship between measurement intensity and reliability is essentially the same as in the classical case of scalar measures. In Sections 6.2 and 6.3, we investigate the extent of agreement between what is expected theoretically and what is observed with simulated and real data.

6.1 | An SB formula for covariance matrix estimation

Let $\Sigma_1, \dots, \Sigma_I$ be a random sample of $p \times p$ covariance matrices, and for $i \in \{1, \dots, I\}$, let $S_{i1}, \dots, S_{iJ_i}$ be sample covariance matrices, each based on m independent and identically distributed (IID) observations $\mathcal{X}_{ij1}, \dots, \mathcal{X}_{ijm}$ from a p-variate normal distribution with covariance matrix Σ_i. These belong to the Hilbert space $\mathcal{H}$ of real symmetric $p \times p$ matrices, equipped with inner product $\langle A, B\rangle = \text{tr}(AB^T)$; the norm induced by this inner product is the Frobenius (entry-wise ℓ_2) norm used in the fMRI example of Section 4. Note that here, unlike in the classical true score model, $T_i \equiv \Sigma_i$ and $\varepsilon_{ij} \equiv S_{ij} - \Sigma_i$ are not independent since ε_{ij} must be such that $S_{ij} = T_i + \varepsilon_{ij}$ is nonnegative definite. But as shown in the Appendix, assumptions (a1) and (a2) of Section 5.2 hold, and consequently,

$$\Delta_\varepsilon(m) \propto \frac{1}{m-1}. \qquad (16)$$

Thus by (15),

$$\frac{\rho_m}{1 - \rho_m} \propto m - 1; \qquad (17)$$

this is almost exactly the classical SB relation (11), but with $m - 1$ in place of m.

6.2 | Log-log plots with simulated data

Suppose that, for a given collection $\Sigma_1, \dots, \Sigma_I$ of $p \times p$ covariance matrices, we repeatedly generate sets of sample covariances as in Section 6.1, but with varying values of m, and obtain a dbICC estimate $\hat{\rho}_m$, based on the ℓ_2 distance, for each m. Then the relation (17) suggests that the points

$$[\log(m-1), \log\{\hat{\rho}_m/(1 - \hat{\rho}_m)\}] \qquad (18)$$

should lie approximately along a line with slope 1. To test this suggestion with simulated data resembling the fMRI data analyzed in Sections 4 and 6.3, we followed the above recipe with

- $I = 25$, $J_i \equiv 2$ and $p = 333$;
- Σ_i ($i = 1, \dots, 25$) taken to be the mean of the two sample covariance matrices from the ith participant's two fMRI scans; and

- a range of m values from 25 to 197, approximately equally spaced on the log scale.

A plot of the resulting points (18) appears in the left panel of Figure 6 (black dots), and the best-fit line through these points has slope 0.997 with standard error 0.010, in agreement with the theoretical slope 1.

Many aspects of the fMRI data reliability analysis in Section 4 are not captured by the above simulation setup. Two of the most prominent disparities are that for the real data, (a) we computed dbICC for correlation, rather than covariance, matrices, and (b) the multivariate observations are autocorrelated rather than independent (see Arbabshirani *et al.*, 2014 and Zhu and Cribben, 2018, regarding the impact of such autocorrelation).

The simulation study was expanded to partially address these discrepancies. Using a standard implementation (Barbosa, 2012) for vector autoregressive models of order 1 (VAR(1); Lütkepohl, 2005), we conducted further simulations in which the jth multivariate time series for the ith individual was given by $x_t^{(ij)} = \phi x_{t-1}^{(ij)} + u_t^{(ij)}$ ($t = 2, \dots, m$), with independent innovations $u_t^{(ij)}$ having zero mean and 333×333 covariance matrix Σ_i. The lag-1 autocorrelation ϕ was set to the values 0.6 and 0.9, which are near the low and high ends of the range of AR(1)-model-based estimates for individual ROIs in our fMRI data. The resulting points (18), with $\hat{\rho}_m$ derived from sample covariance matrices, are displayed in the left panel of Figure 6. The right panel is analogous, but here $\hat{\rho}_m$ is derived from sample correlation matrices. A comparison of the two panels indicates that, for given autocorrelation settings, both the estimated SNR $\hat{\rho}_m/(1 - \hat{\rho}_m)$ and its dependence on m are very similar for covariance versus correlation matrix estimation. Autocorrelation is seen to reduce reliability and thus to shift the SNR markedly downward. Moreover, autocorrelation seems to attenuate the linear relationship between m and SNR: whereas in the IID setting, the slope is 1.018 for the sample correlation matrix, again very close to the theoretical value 1, the slopes are smaller with autocorrelation 0.6 (0.986 for covariance, 0.960 for correlation) and even smaller for autocorrelation 0.9 (0.736 for covariance, 0.687 for correlation). In Web Appendix C, we present plots that are analogous to Figure 6, but based on the ℓ_1 and $\sqrt{1 - r}$ distances, and we report the intercepts and slopes of the best-fit lines for all cases.

6.3 | Reliability based on subsets of the fMRI time series

Next, we constructed log-log plots as above but based on subsets of the real fMRI time series of Section 4 rather than on simulated data. For values of m ranging from 25 to the full time series length 197, we took the middle m observations

from each of the fMRI time series, and thus computed correlation matrices R_{ij} ($i = 1, \dots, 25; j = 1, 2$) using the same three sets of ROIs as in Section 4: all 333 ROIs proposed by Gordon *et al.* (2016), the default mode network, and the visual network. Log-log plots for the resulting dbICC values $\hat{\rho}_m$ appear in the right panel of Figure 6. For smaller m, these plots are quite nonlinear and distinct from each other, but for $m > 100$, they each appear to stabilize with a linear pattern that is roughly parallel to the best-fit line for the simulations with lag-1 autocorrelation .9.

This degree of agreement with the simulation results of Section 6.2 is probably as much as can be expected, given the significant discrepancies between the settings of the simulated- and real-data analyses, which include the following: (a) The simulations for different m are independent, whereas with the real data, for increasing m, we consider a nested sequence of increasingly large subsets of the same time series. (b) The real time series may not be multivariate normal and presumably have more complex patterns of autocorrelations and cross-correlations than the simulated data.

At any rate, it seems clear that the theoretical log-log plot slope of 1 cannot be expected to characterize the reliability improvement attainable via longer fMRI time series. Our results offer hope that a slope around 0.7 might be attained, but at least two further caveats are in order. One is that we cannot extrapolate beyond $m = 197$, the full time series length for our data. A second, subtler caveat concerns the true score model (1), in the specific form outlined in Section 6.1. That model assumes that for each i, the two sample covariance matrices S_{i1}, S_{i2} are estimates of a common true covariance Σ_i. But if, in fact, the underlying covariance matrix differs between the two fMRI scans for at least some of the participants, this is an additional source of within-subject distance that is not removed by increasing the time series length m, and thus, $\log[\hat{\rho}_m/(1 - \hat{\rho}_m)]$ may tend to level off rather than increasing linearly with $\log(m - 1)$. In summary, while longer fMRI scans might make correlation matrices more reliable as measures of functional connectivity, the improvement would likely be less dramatic than the results reported here might lead us to expect.

7 | FURTHER APPLICATION AND EXTENSION OF THE SB FORMULA

Log-log plots like those in Figure 6 are a broadly applicable tool for examining the relationship between measurement intensity m and reliability. As discussed in Web Appendix D, for penalized spline smoothing (example (E2) of Section 5.1), $\Delta_\varepsilon(m) \propto m^{-8/9}$. Thus, arguing as in Section 6.2, a linear model fit to the points $[\log(m), \log\{\hat{\rho}_m/(1 - \hat{\rho}_m)\}]$ should have slope $\frac{8}{9}$, a prediction that is borne out with simulated data.
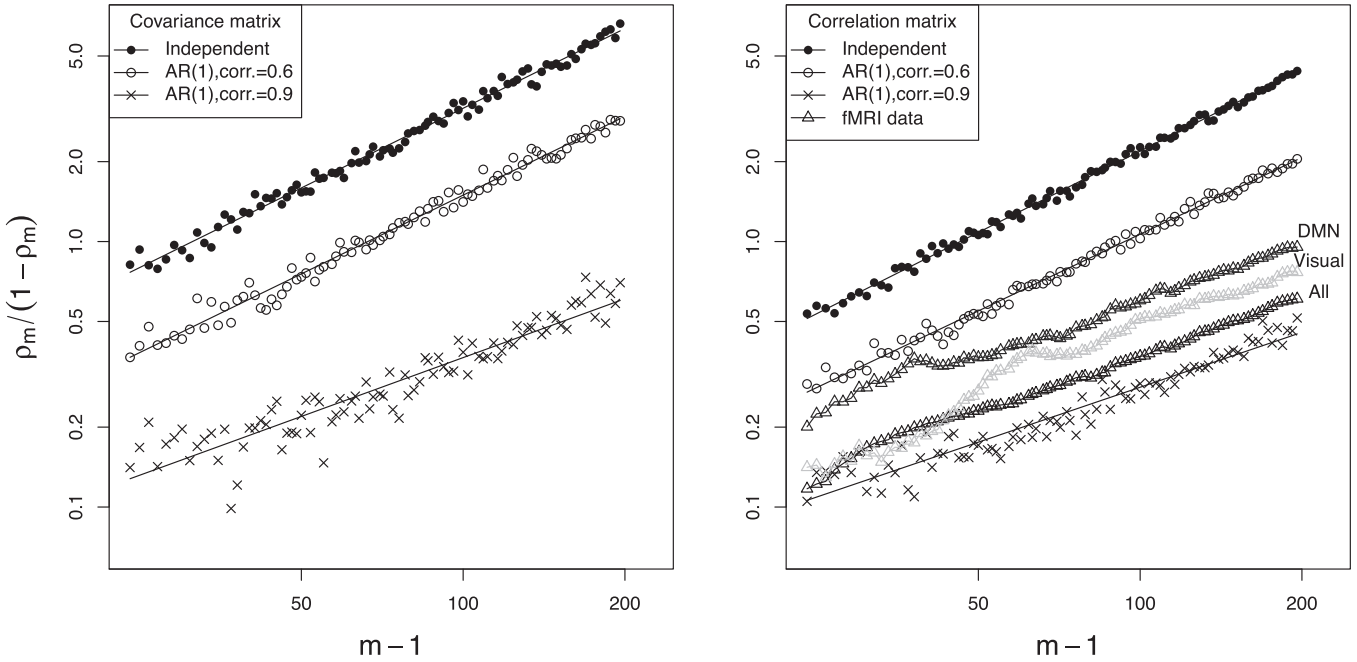
FIGURE 6 *Left:* Effect of measurement intensity on SNR $\frac{\rho}{1-\rho}$ for covariance matrix estimation with simulated data. Both axes are plotted on the log scale since, as explained at (18), this is expected to yield a linear relation with slope 1 for independent observations. *Right:* Simulation results for correlation matrix estimation, along with results based on subsets of the fMRI time series

Some distances, such as the dynamic time warping distance between signatures considered in Web Appendix E, do not arise from the true score model (1), even in the generalized (Hilbert space-valued) form of Section 5.2. Whether or not the true score model applies, the dbICC (3) satisfies

$$\frac{\rho}{1-\rho} = \frac{\mathrm{MSD}_b - \mathrm{MSD}_w}{\mathrm{MSD}_w}. \tag{19}$$

The key to the derivation of (15) is simply that, by (12)-(14),

(i) $\mathrm{MSD}_w = \mathrm{MSD}_w(m) = \Delta_\varepsilon(m)$,
(ii) $\mathrm{MSD}_b - \mathrm{MSD}_w = \Delta_T$, which does not depend on m.

The same argument works more generally (ie, not only in Hilbert spaces): as long as MSD_w can be written as a function of m whereas $\mathrm{MSD}_b - \mathrm{MSD}_w$ does not change with m, it follows from (19) that

$$\frac{\rho_m}{1-\rho_m} \propto \frac{1}{\mathrm{MSD}_w(m)}, \tag{20}$$

generalizing (15), which is itself a generalization of (11).

Log-log plots might be used in this more general setting to estimate the effect of measurement intensity m on ρ_m, as opposed to confirming a theoretical relationship. By (20), if it is expected that $\mathrm{MSD}_w(m) \propto m^{-\beta}$ for some unknown β, then we can regress values of $\log \frac{\hat{\rho}_m}{1-\hat{\rho}_m}$ on the corresponding values of $\log(m)$, and the resulting slope serves as an estimate of β.

A similar approach is used to estimate the Hurst exponent of a long memory process (Beran, 1994).

8 | DISCUSSION

In this paper, we have redefined the intraclass correlation coefficient in terms of distances, and thereby extended this reliability index to arbitrary data objects for which a distance is defined. The proposed distance-based ICC leads to two extensions of the SB formula, namely, (15) for Hilbert space-valued data including covariance matrices, and (20) for more general data objects.

In an early paper on extending the ICC to multivariate data, Fleiss (1966) wrote that a classical (univariate) ICC value less than about 0.70 "is, for most purposes, taken to indicate insufficient reliability." The much lower dbICC values that we report for functional connectivity data, along with similar results reported by others (eg, Shou *et al.*, 2013), are a sobering indication that in some cases, as technology has advanced, the reliability of complex new measures has retreated. This might help to explain the recently-much-discussed difficulties surrounding scientific reproducibility, a desideratum that is closely related to reliability (Yu, 2013).

While our presentation has focused on test-retest data, the dbICC might also be applied to assess the reliability of results obtained by algorithms, such as bootstrapping, which have a stochastic component (cf. Philipp *et al.*, 2018).

While we have developed a distance-based analog of the *intra*class correlation coefficient, the distance correlation of Székely *et al.* (2007) is comparable to *inter*class correlation coefficients. Extending ideas from distance correlation research to the intraclass setting may be an interesting avenue for future work.

ACKNOWLEDGMENTS

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available in the Supporting Information. These data were derived from the public-domain NYU CSC TestRetest resource at http://www.nitrc.org/projects/nyu_trt.

ORCID

Meng Xu https://orcid.org/0000-0002-0544-3842
Philip T. Reiss https://orcid.org/0000-0002-8491-7080
Ivor Cribben https://orcid.org/0000-0001-6548-581X

REFERENCES

Alonso, A., Laenen, A., Molenberghs, G., Geys, H. and Vangeneugden, T. (2010) A unified approach to multi-item reliability. *Biometrics*, 66, 1061–1068.

Arbabshirani, M.R., Damaraju, E., Phlypo, R., Plis, S., Allen, E., Ma, S. et al. (2014) Impact of autocorrelation on functional connectivity. *NeuroImage*, 102, 294–308.

Atenafu, E.G., Hamid, J.S., To, T., Willan, A.R., Feldman, B.M. and Beyene, J. (2012) Bias-corrected estimator for intraclass correlation coefficient in the balanced one-way random effects model. *BMC Medical Research Methodology*, 12, 126.

Barbosa, S.M. (2012) *mAr: Multivariate AutoRegressive Analysis*. R package version 1.1-2.

Beran, J. (1994) *Statistics for Long-Memory Processes*. Boca Raton, FL: CRC Press.

Brown, W. (1910) Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.

Choe, A.S., Nebel, M.B., Barber, A.D., Cohen, J.R., Xu, Y., Pekar, J.J., Caffo, B. and Lindquist, M.A. (2017) Comparing test-retest reliability of dynamic functional connectivity methods. *NeuroImage*, 158, 155–175.

Cranford, J.A., Shrout, P.E., Iida, M., Rafaeli, E., Yip, T. and Bolger, N. (2006) A procedure for evaluating sensitivity to within-person change: can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, 32, 917–929.

Fleiss, J.L. (1966) Assessing the accuracy of multivariate observations. *Journal of the American Statistical Association*, 61, 403–412.

Fleiss, J.L. (1986) *Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons.

Fujikoshi, Y., Ulyanov, V.V. and Shimizu, R. (2010) *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Hoboken, NJ: John Wiley & Sons.

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M. and Petersen, S.E. (2016) Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral Cortex*, 26, 288–303.

Lord, F.M. and Novick, M.R. (1968) *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lütkepohl, H. (2005) *New Introduction to Multiple Time Series Analysis*. New York: Springer Science & Business Media.

Mair, P. (2018) *Modern Psychometrics with R*. Cham, Switzerland: Springer.

McArdle, B.H. and Anderson, M.J. (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82, 290–297.

Mielke, P.W. and Berry, K.J. (2007) *Permutation Methods: A Distance Function Approach*. New York: Springer.

Noble, S., Scheinost, D. and Constable, R.T. (2019) A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. *Neuroimage*, 203, 116157.

Philipp, M., Rusch, T., Hornik, K. and Strobl, C. (2018) Measuring the stability of results from supervised statistical learning. *Journal of Computational and Graphical Statistics*, 27, 685–700.

R Core Team (2019) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A. and Shulman, G.L. (2001) A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98, 676–682.

Reiss, P.T., Stevens, M.H.H., Shehzad, Z., Petkova, E. and Milham, M.P. (2010) On distance-based permutation tests for between-group comparisons. *Biometrics*, 66, 636–643.

Shehzad, Z., Kelly, A.C., Reiss, P.T., Gee, D.G., Gotimer, K., Uddin, L.Q. et al. (2009) The resting brain: unconstrained yet reliable. *Cerebral Cortex*, 19, 2209–2229.

Shou, H., Eloyan, A., Lee, S., Zipunnikov, V., Crainiceanu, A., Nebel, M. et al. (2013) Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2). *Cognitive, Affective, & Behavioral Neuroscience*, 13, 714–724.

Somandepalli, K., Kelly, C., Reiss, P.T., Zuo, X.-N., Craddock, R.C., Yan, C.-G., Petkova, E., Castellanos, F.X., Milham, M.P. and Di Martino, A. (2015) Short-term test–retest reliability of resting state fMRI metrics in children with and without attention-deficit/hyperactivity disorder. *Developmental Cognitive Neuroscience*, 15, 83–93.

Spearman, C. (1910) Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295.

Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35, 2769–2794.

Yan, C.-G., Craddock, R.C., Zuo, X.-N., Zang, Y.-F. and Milham, M.P. (2013) Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. *NeuroImage*, 80, 246–262.

Yu, B. (2013) Stability. *Bernoulli*, 19, 1484–1500.

Zhu, Y. and Cribben, I. (2018) Sparse graphical models for functional connectivity networks: best methods and the autocorrelation issue. *Brain Connectivity*, 8, 139–165.

SUPPORTING INFORMATION

Web Appendix A, referenced in Section 1, Web Appendix B, referenced in Section 4, Web Appendix C, referenced in Section 6.2, and Web Appendices D and E, referenced in Section 7, along with a brief guide to the R code, are available with this paper at the Biometrics website on Wiley Online Library. A package for R (R Core Team, 2019) implementing the methods of this paper is available at https://github.com/wtagr/dbicc.

APPENDIX

fMRI data description and preprocessing

The resting-state fMRI data set, downloaded from http://www.nitrc.org/projects/nyu_trt, includes 25 participants (mean age 29.44 ± 8.64, 10 males) scanned at New York University. A Siemens Allegra 3.0-Tesla scanner was used to obtain three resting-state scans for each participant, although for this analysis, we considered only the second and third scans, which were less than 1 hour apart. Each scan consisted of 197 contiguous EPI functional volumes with time repetition (TR) = 2000 ms; time echo (TE) = 25 ms; flip angle (FA) = 90°; 39 number of slices, matrix = 64 × 64; field of view (FOV) = 192 mm; voxel size $3 \times 3 \times 3$ mm^3. During each scan, the participants were asked to relax and remain still with eyes open. For spatial normalization and localization, a high-resolution T1-weighted magnetization prepared gradient echo sequence was obtained (MPRAGE, TR = 2500 ms; TE = 4.35 ms; inversion time = 900 ms; FA = 8°, number of slices = 176; FOV = 256 mm).

The data were preprocessed using the FSL (http://www.fmrib.ox.ac.uk) and AFNI (http://afni.nimh.nih.gov/afni) software packages. The images were (a) motion corrected using FSL's mcflirt (rigid body transform; cost function normalized correlation; reference volume the middle volume) and then (b) normalized into the Montreal Neurological Institute space using FSL's flirt (affine transform; cost function mutual information). (c) FSL's fast was then used to obtain a probabilistic segmentation of the brain to acquire white matter and cerebrospinal fluid (CSF) probabilistic maps, thresholded at 0.99. (d) AFNI's 3dDetrend was then used to remove the nuisance signals, namely, the six motion parameters, white matter and CSF signals, and the global signal. (e) Finally, using FSL's fslmaths, the volumes were spatially smoothed using a Gaussian kernel with FWHM = 6 mm.

The ROIs for our connectivity analysis are derived from the work of Gordon *et al.* (2016), who parcellated the cortical surface into 333 areas within which homogeneous connectivity patterns are observed. Time courses for these 333 ROIs were obtained for each subject by averaging over all of the voxels within each region. Each regional time course was then detrended and standardized to unit variance, and then we applied a fourth-order Butterworth filter with passband 0.01-0.10 Hertz.

(a1), (a2), and $\Delta_\varepsilon(m)$ for sample covariance matrices

Sample covariance matrices of multivariate normal samples are a special case of the true score model of Section 5.2 in which, for each i, $T_i = \Sigma_i$, a $p \times p$ covariance matrix, and for each i, j,

$$\varepsilon_{ij} = S_{ij} - \Sigma_i, \tag{A.1}$$

where S_{ij} is the sample covariance matrix of an IID random sample $\mathcal{X}_{ij1}, \ldots, \mathcal{X}_{ijm} \sim N_p(\mathbf{0}, \Sigma_i)$. Here, we verify assumptions (a1) and (a2) of Section 5.2 for this case, and derive expression (16) for $\Delta_\varepsilon(m)$.

By (A.1), $\varepsilon_{i_1 j_1}, \varepsilon_{i_2 j_2}$ in (13) are independent mean-zero matrices, implying that

$$\Delta_\varepsilon(m) = E\left[\text{tr}\{(\varepsilon_{i_1 j_1} - \varepsilon_{i_2 j_2})^2\} \right]$$
$$= E[\text{tr}(\varepsilon_{i_1 j_1}^2)] + E[\text{tr}(\varepsilon_{i_2 j_2}^2)] - 2E[\text{tr}(\varepsilon_{i_1 j_1} \varepsilon_{i_2 j_2})].$$

For $i_1 \neq i_2$, $E[\text{tr}(\varepsilon_{i_1 j_1} \varepsilon_{i_2 j_2})] = 0$ since $\varepsilon_{i_1 j_1}, \varepsilon_{i_2 j_2}$ are independent mean-zero matrices. On the other hand, if $i_1 = i_2 = i$, then $\varepsilon_{i_1 j_1}, \varepsilon_{i_2 j_2}$ are independent and of mean zero, conditionally on Σ_i, and thus again

$$E[\text{tr}(\varepsilon_{i_1 j_1} \varepsilon_{i_2 j_2})] = E[E\{\text{tr}(\varepsilon_{i_1 j_1} \varepsilon_{i_2 j_2})|\Sigma_i\}] = 0.$$

Hence, the expectation defining $\Delta_\varepsilon(m)$ does not depend on whether or not $i_1 = i_2$, ie, (a1) holds; and

$$\Delta_\varepsilon(m) = 2E[\text{tr}(\varepsilon_{ij}^2)], \tag{A.2}$$

for ε_{ij} as in (A.1).

For (a2), it suffices to show that $E[\text{tr}\{\Sigma_{i1}(\varepsilon_{i_1 j_1} - \varepsilon_{i_2 j_2})\}] = 0$. This follows since

$$E[\text{tr}(\Sigma_{i_1} \varepsilon_{i_1 j_1})] = E[E\{\text{tr}(\Sigma_{i1} \varepsilon_{i_1 j_1})|\Sigma_{i_1}\}] = 0,$$

while $E[\text{tr}(\Sigma_{i1} \varepsilon_{i_2 j_2})] = 0$ since $\varepsilon_{i_2 j_2}$ is independent of Σ_{i1} and of mean zero.

By a standard result in multivariate analysis, conditionally on $\mathbf{\Sigma}_i$, $(m-1)\mathbf{S}_{ij}$ has a Wishart($\mathbf{\Sigma}_i$) distribution with $m-1$ degrees of freedom; thus by Theorem 2.2.6 of Fujikoshi *et al.* (2010),

$$E[\mathrm{tr}(\mathbf{S}_{ij}^2)|\mathbf{\Sigma}_i] = \frac{1}{m-1}\left[(\mathrm{tr}\mathbf{\Sigma}_i)^2 + m\mathrm{tr}(\mathbf{\Sigma}_i^2)\right] \quad \text{and}$$

$$E[\mathrm{tr}(\mathbf{S}_{ij}\mathbf{\Sigma}_i)|\mathbf{\Sigma}_i] = \mathrm{tr}(\mathbf{\Sigma}_i^2).$$

These results lead to

$$E[\mathrm{tr}(\varepsilon_{ij}^2)|\mathbf{\Sigma}_i] = E[\mathrm{tr}\{(\mathbf{S}_{ij} - \mathbf{\Sigma}_i)^2\}|\mathbf{\Sigma}_i]$$

$$= \frac{1}{m-1}[(\mathrm{tr}\mathbf{\Sigma}_i)^2 + \mathrm{tr}(\mathbf{\Sigma}_i^2)].$$

Combining this with (A.2) gives

$$\Delta_\varepsilon(m) = \frac{2}{m-1}E[(\mathrm{tr}\mathbf{\Sigma}_i)^2 + \mathrm{tr}(\mathbf{\Sigma}_i^2)],$$

where the expectation is with respect to the distribution of the true covariance matrices $\mathbf{\Sigma}_i$. This confirms (16).