Behav Res (2016) 48:994–1003
DOI 10.3758/s13428-015-0623-y

Choosing the best index for the average score intraclass correlation coefficient

Gwowen Shieh [1]

Published online: 17 July 2015

Abstract The intraclass correlation coefficient (ICC)(2) index from a one-way random effects model is widely used to describe the reliability of mean ratings in behavioral, educational, and psychological research. Despite its apparent utility, the essential property of ICC(2) as a point estimator of the average score intraclass correlation coefficient is seldom mentioned. This article considers several potential measures and compares their performance with ICC(2). Analytical derivations and numerical examinations are presented to assess the bias and mean square error of the alternative estimators. The results suggest that more advantageous indices can be recommended over ICC(2) for their theoretical implication and computational ease.

Keywords Multilevel modeling · Random effect · Reliability

The most fundamental phenomenon about the hierarchical nature of individual and group influences in multilevel research is that measurements on individuals (e.g., employee, student, patient) within the same group (e.g., organization, classroom, clinic) are presumably more similar than measurements on individuals in different groups. Accordingly, various forms of the intraclass correlation coefficient (ICC) have been proposed to represent the reliability or degree of resemblance among cluster members. Essentially, they can be interpreted as the proportion of the total variance of the response that is accounted for by the clustering or group cohesion. However, different conceptual frameworks and modeling formulations of a multilevel study ultimately lead to distinct and unique definition of ICCs. Comprehensive reviews and general guidelines were provided in Bartko (1976), McGraw and Wong (1996), and Shrout and Fleiss (1979) for selecting the appropriate model and ICC as an interrater reliability measure in one-way random effects and two-way random effects or mixed effects models. Moreover, definitional issues and methodological appraisals concerning ICC, interrater reliability, and interrater agreement can be found in Bliese (2000), James (1982), LeBreton et al. (2003), LeBreton and Senter (2008), and the references therein.

To assess the magnitude of similarity or interrelation of hierarchical data, the ICC(1) and ICC(2) indices based on the one-way random effects model are the two most frequently adopted reliability measures for the single score and average score ICCs, respectively, within the context of multilevel modeling. Specifically, the well-established single score and average score ICCs ρ and ρ^* are defined as

$$\rho = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\varepsilon^2},$$

and

$$\rho^* = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\varepsilon^2/K},$$

respectively, where σ_γ^2 represents the between-group variance, σ_ε^2 is the within-group variance, and K is the group size. The two definitions of ICCs reveal that the average score ρ^* is always greater in magnitude than the single score counterpart ρ and the magnitude of ρ^* is greatly influenced by the group size.

✉ Gwowen Shieh
gwshieh@mail.nctu.edu.tw

[1] Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan 30010, Republic of China

The most commonly used estimators of ρ and ρ* are given by

$$\mathrm{ICC}(1) = \frac{MSB - MSW}{MSB + (K-1)MSW} = \frac{F^* - 1}{F^* + K - 1},$$

and

$$\mathrm{ICC}(2) = \frac{MSB - MSW}{MSB} = 1 - \frac{1}{F^*},$$

where *MSB* is the between-group mean square, *MSW* is the within-group mean square, and $F^* = MSB/MSW$ calculated from the one-way random effects model. This index ICC(2) follows the notation of Bartko (1976), Bliese (2000), and James (1982). However, it has also been referred to as ICC(*k*) in McGraw and Wong (1996) and as ICC(1, *k*) in Shrout and Fleiss (1979). In general, ICC(1) is an estimate of effect size indicating the extent to which individual ratings are attributable to group membership, whereas ICC(2) estimates the reliability of mean ratings furnished by a group of judges.

The respective magnitudes of ICC(1) and ICC(2) allow researchers to appraise the level of observed variance of single score and average score that is affected by clustering. For example, in a two-level analysis of the influence of classroom climate perceptions on individual students' levels of academic achievement, a ICC(1) value of 0.2 indicates that 20 % of the observed variance in students' achievement scores is due to systematic between-classroom differences compared to the total variance in achievement scores. In contrast, a value of ICC(2) = 0.8 represents that 80 % of the observed total variance in classroom average scores occurring at the classroom level. Consequently, the use and interpretation of ICC(1) and ICC(2) are appropriate if a researcher is interested in drawing inferences concerning the reliability of single score and average score, respectively.

To further illustrate the fundamental differences between the two indices, it is instructive to consider the ICC(1) value of 0.10 yields the ICC(2) of 0.7, 0.8, and 0.9 for the group size of 21, 36, and 81, respectively. The review of climate studies in James (1982) showed that ICC(1) values range from 0.0 to 0.5 with a median of approximately 0.12. Moreover, Hedges and Hedberg (2007) reported that the resulting ICC(1) values for a variety of school performance studies are generally in the range of 0.10 to 0.25. In contrast, a ICC(2) value of 0.7 has been widely used as the minimum acceptable level of reliability for psychological measures. However, Lance, Butts, and Michels (2006) noted that many researchers did not provide adequate justification for the appropriateness of the commonly used cut point of 0.7. Consequently, it should not be treated as a universal standard.

Within the context of one-way random effects modeling, it follows from the standard results that $E[MSB] = K\sigma_\gamma^2 + \sigma_\varepsilon^2$ and $E[MSW] = \sigma_\varepsilon^2$ (McGraw & Wong, 1996, Table 3). Hence, $(MSB - MSW)/K$ and *MSW* are unbiased estimators of σ_γ^2 and σ_ε^2, respectively. From the viewpoint of estimation principle, ICC(1), introduced by Fisher (1938), is obtained by substituting the variance components in population single score ICC with corresponding unbiased estimators. Although this natural modification is intuitive and heuristic, ICC(1) is not an unbiased estimator of the corresponding individual rating ICC. The interested reader is referred to Searle, Casella, and McCulloch (1992) for further technical details of various methods for estimating the within-group and between-group variances of random effects models. Note that Olkin and Pratt (1958) have derived the minimum variance unbiased estimator of the single score ICC, but its use has been impeded by the lack of a closed form expression. The corresponding computation requires a special purpose computer program; see, for example, Donoghue and Collins (1990). Moreover, Donner (1986) and Harris and Burch (2000) presented extensive discussions of compelling alternatives and associated properties for estimating the individual rating ICC.

In addition to the theoretical developments in statistical literature, Bliese and Halverson (1998) suggested the corrected eta-squared formula as a modification of sample eta-squared estimator to provide more accurate estimates of the single score ICC. With the emphasis on the analysis of group-level properties in organizational research, the empirical investigation of Bliese and Halverson (1998) focused on the behavior of sample eta-squared estimator. The numerical results showed that sample eta-squared is a positively biased estimate of the individual rating ICC and the performance varies with group size and the magnitude of population intraclass correlation. However, they did not examine the inherent properties of the corrected eta-squared formula. Shieh (2012) recently showed that the corrected eta-squared estimator described in Bliese and Halverson (1998) is identical to the maximum likelihood estimator and presented an extensive comparison between their truncated versions for negative values. The modification of corrected eta-squared estimator performs better when the underlying population single score ICC is small. Conversely, the adjusted ICC(1) has a clear advantage for medium and large magnitudes of population individual rating ICC. Thus, the existing findings have concluded that although ICC(1) is the best known, it may not always be the best choice.

Unlike the prevalent attention and investigation of ICC(1) and related indices for the analysis of multilevel questions, the theoretical property and intrinsic appropriateness of ICC(2) for the estimation of the average score ICC have been given insufficient consideration in the literature. Basically, the average score ICC is a function of the individual rating ICC through the Spearman-Brown prophesy formula (Brown, 1910; Spearman, 1910). It also can be readily established that the formulation of ICC(2) is equivalent to the Spearman-

Brown equation by replacing the population single score ICC with ICC(1). Note that the desirable estimation property of an individual rating ICC index for the population single score ICC does not naturally extended to the corresponding Spearman-Brown counterpart for the estimation of the associated population average score ICC. Despite the direct connection between the individual rating and average score ICCs, ICC(2) not only has a unique interpretation as a reliability index of group mean rating, but also possesses completely different properties from ICC(1). The existing findings associated with ICC(1) are arguably not suitable to demonstrate the explicit performance of ICC(2). More importantly, the estimation problem of the mean rating ICC should be duly recognized and it requires a unified and rigorous treatment to clarify the stochastic behavior of feasible solutions.

The continual use of ICC(2) as the standard average score ICC index without identification of the essential limitations may not facilitate a better interpretation and application of research findings. For the ultimate aim of selecting the most appropriate methodology, it is vital to ensure that the contrasting properties of ICC(2) and viable alternatives are thoroughly explicated. The present article purports to contribute to the literature on choosing the best index of the average score ICC within the framework of one-way random effects model. First, a simplified expression is presented to synthesize the essential attributes of the single score ICC estimators in Gleason (1997) and Harris and Burch (2000). Then the Spearman-Brown formula is applied to obtain a useful class of estimators of the average score ICC. Second, in order to judge the merits of various measures from the point estimation perspective, explicit analytic forms of the bias and mean square error (MSE) are derived for the considered mean rating ICC indices. Accordingly, the optimal estimators under bias and MSE considerations are identified. Third, numerical appraisals are performed to illustrate the relative performance of the renowned ICC(2) and several desirable measures within the suggested family of average score ICC estimators. A discussion of potential implications of the findings for both theoretical development and practical use in reliability study is also presented.

Estimation of the average score intraclass correlation coefficient

Within the context of multilevel analysis, a widely used design is the one-way random effects model

$$Y_{ij} = \mu + \gamma_i + \varepsilon_{ij}, i = 1, \ldots, N; j = 1, \ldots, K, \quad (1)$$

where Y_{ij} is the *jth* individual measurement within group i, μ is the grand mean, and γ_i and ε_{ij} are independent random variables with $\gamma_i \sim N(0, \sigma_\gamma^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. The variance of Y_{ij} is then given by $\sigma_\gamma^2+\sigma_\varepsilon^2$. Accordingly, the single score ICC $\rho=\sigma_\gamma^2/(\sigma_\gamma^2+\sigma_\varepsilon^2)$ is the simple correlation coefficient $Corr(Y_{i1}, Y_{i2})$ between any two observations, Y_{i1} and Y_{i2}, in the same group. The single score ICC ρ may also be interpreted as the proportion of the total variance of individual response that is accounted for by the clustering or group cohesion. In contrast, the simple correlation coefficient $Corr\left(\overline{Y}_{i(1)}, \overline{Y}_{i(2)}\right)$ between any two sets of mean measurements, $\overline{Y}_{i(1)} = \sum_{j=1}^{K} Y_{ij}/K$ and $\overline{Y}_{i(2)} = \sum_{j=K+1}^{2K} Y_{ij}/K$, from the same group is defined as the average score ICC $\rho^*=\sigma_\gamma^2/(\sigma_\gamma^2+\sigma_\varepsilon^2/K)$. It also represents the proportion of the total variance of mean ratings for a group of K judges that is accounted for by the grouping or cluster membership. The prominent coefficient ρ^* can also be written in the form of Spearman-Brown prediction formula $\rho^* = \Psi(\rho)$ where

$$\Psi(\rho) = \frac{K\rho}{1 + (K-1)\rho}. \quad (2)$$

It is straightforward to show that the ICC(2) index is basically the Spearman-Brown formula applied to ICC(1) or ICC(2) = Ψ{ICC(1)} for any value $K > 1$. This particular result was also noted in James (1982) and is more precise than the asymptotic equivalence between ICC(2) and Ψ{ICC(1)} demonstrated in Bliese (1998).

The simple notion of substituting the individual rating index ICC(1) into the Spearman-Brown prophesy formula to attain the average score measure ICC(2) suggests a integrated approach to utilizing the existing procedures for the estimation of individual rating ICC to the estimation of average score ICC. Because the complexity of a functional form may result in limited acceptance for practical use, the most appealing feature of a practically useful index is its computational simplicity. To this end, only the estimators with a convenient analytical form are considered here. The following unified expression is presented to accommodate and simplify the diverse individual rating estimators in Gleason (1997) and Harris and Burch (2000):

$$\hat{\rho}(c) = \frac{F^*-c}{F^* + cK-c}, \quad (3)$$

where c is a constant. Clearly, because $\hat{\rho}(1) = (F^* - 1)/(F^* + K - 1) = \text{ICC}(1)$, $\hat{\rho}(c)$ includes ICC(1) as a special case when $c = 1$. More importantly, with the application of Spearman-Brown equation to $\hat{\rho}(c)$, a simple closed-form expression is acquired for the suggested class of average score estimators:

$$\hat{\rho}^*(c) = \Psi\left\{\hat{\rho}(c)\right\} = 1-\frac{c}{F^*}. \quad (4)$$

As expected, it is seen that ICC(2) = $\hat{\rho}^*(1)$ is a significant member of $\hat{\rho}^*(c)$. Note that ICC(1) is obtained by replacing variance parameters in population ICC ρ with corresponding unbiased estimators. Because it is often called the ANOVA

estimator, for ease of illustration, a specific notation, ICC(2) = $\hat{\rho}^*_{AV} = \hat{\rho}^*(c_{AV})$ with $c_{AV} = 1$, is given to denote the particular instance. The other notable choices $\{\hat{\rho}^*_{MO},\ \hat{\rho}^*_{ME},\ \hat{\rho}^*_{EF},\ \hat{\rho}^*_{ML}\}$ *of* $\hat{\rho}^*(c)$ with $c = \{c_{MO}, c_{ME}, c_{EF}, c_{ML}\}$ considered in Gleason (1997) and Harris and Burch (2000) are given as follows:

$c_{MO} = \{N(N-3)(K-1)\}/\{(N-1)[N(K-1)+2]\}$ is the mode of the F distribution $F\{N-1, N(K-1)\}$ with $N-1$ and $N(K-1)$ degrees of freedom;
$c_{ME} = F_{(N-1),\ N(K-1),\ 0.5}$, where $F_{(N-1),\ N(K-1),\ 0.5}$ is the median of the F distribution $F\{N-1, N(K-1)\}$;
$c_{EF} = \{N(K-1)\}/\{N(K-1)-2\}$ is the expected value of the F distribution $F\{N-1, N(K-1)\}$;
$c_{ML} = N/(N-1)$ corresponds to the application of the maximum likelihood estimators of σ_γ^2 and σ_ε^2.

It should be emphasized that the proposed class of average score estimators not only possesses the major advantage in the ease of application, but also facilitates exact theoretical justification of the associated properties presented in Appendix. It follows from Eqs. A3 and A5 that $\hat{\rho}^*_{UB} = \hat{\rho}^*(c_{UB})$ and $\hat{\rho}^*_{MS} = \hat{\rho}^*(c_{MS})$ are the best unbiased and the best MSE estimators within the considered class of indices, respectively, where

$$c_{UB} = \frac{N-3}{N-1} \quad \text{and} \quad c_{\mathrm{MS}} = \frac{N(N-5)(K-1)}{N-1\{N(K-1)+2\}}.$$

Unlike the computational demand of the minimum variance unbiased estimator of individual rating ICC, these two optimal indices of average score ICC are considerably convenient for practical use.

An immediate observation from the estimation properties and optimal solutions is that the conventional ICC(2) is suboptimal under both the bias and MSE principles, except for the special situation of $\rho^* = 1$. Specifically, the corresponding bias and MSE of ICC(2) are

$$Bias\{\mathrm{ICC}(2)\} = \frac{-2(1-\rho^*)}{N-3} \quad \text{and} \quad MSE\{\mathrm{ICC}(2)\} = (1-\rho^*)^2 M_1, \tag{5}$$

respectively, where

$$M_1 = 1 - \frac{2(N-1)}{(N-3)} + \frac{(N-1)^2\{N(K-1)+2\}}{N(N-5)(N-3)(K-1)}.$$

This implies that ICC(2) is generally a negatively biased estimator of ρ^*, and the absolute bias and MSE become decreasing as the parameter ICC increases for fixed values of N and K. Conversely, the dominant estimators $\hat{\rho}^*_{UB}$ and $\hat{\rho}^*_{MS}$ provide improvement over ICC(2) against the bias and MSE criteria, respectively. It is of both practical value and theoretical interest to further appraise the similarities and differences between the prescribed measures. But due to the complex nature of the resulting bias and MSE, a complete analytical treatment is not feasible. In order to present a comprehensive explication for the relative merits of different indices, a detailed numerical study is conducted next to explore their estimation behavior.

Numerical illustrations

For the purpose of delineating the essential features of the average score ICC indices, an empirical investigation was performed under a wide range of model configurations. The bias and MSE calculations for the considered estimators require complete specifications of the number of groups, N, the number of judges in each group, K, and the underlying population individual rating ICC, ρ. The numerical computations are systematically conducted and accomplished by fixing all but one of the three decisive attributes and varying a single attribute in the assessment. More importantly, the actual bias and MSE were obtained by one-dimensional numerical integration with respect to an F probability distribution function. The numerical integration is theoretically exact provided that the auxiliary function can be evaluated exactly.

Specifically, two different values (10 and 50) are considered for the number of groups and the number of judges and it leads to four combined scenarios of (N, K) = (10, 10), (10, 50), (50, 10), and (50, 50). It is clear from the Spearman-Brown equation that $\rho^* = \Psi(\rho) = K\rho/\{1 + (K-1)\rho\}$ is a one-to-one function of ρ, and equivalently, $\rho = \rho^*/\{K - (K-1)\rho^*\}$ for a fixed value $K > 1$. For ease of exposition, the values of ρ are chosen so that the resulting ρ^* = 0 to 0.90 with an increment of 0.1 and 0.99. These combinations of model configurations are selected to cover a wide extent of characteristics that are likely to occur in multilevel applications. Overall, the actual performance of bias and MSE of the seven estimators $\{\hat{\rho}^*_{MS},\ \hat{\rho}^*_{MO},\ \hat{\rho}^*_{UB},\ \hat{\rho}^*_{ME},\ \mathrm{ICC}(2),\ \hat{\rho}^*_{EF},\ \hat{\rho}^*_{ML}\}$ are computed for ten different magnitudes of ρ^* for each of the four joined model configurations of two numbers of groups and two group sizes. The bias and MSE results for the four combinations of N and K are summarized in Tables 1, 2, 3 and 4 and Tables 5, 6, 7 and 8, respectively.

To provide a concrete illustration, the relative merits between different estimators are represented by the relative absolute bias and the relative MSE, using ICC(2) as a convenient benchmark. It can be shown from the

Table 1 The bias of average score intraclass correlation coefficient indices for $N = 10$ and $K = 10$

$\hat{\rho}^*$	$\hat{\rho}^*_{MS}$	$\hat{\rho}^*_{MO}$	$\hat{\rho}^*_{UB}$	$\hat{\rho}^*_{ME}$	$\hat{\rho}^*_{AV}$	$\hat{\rho}^*_{EF}$	$\hat{\rho}^*_{ML}$
c	0.5435	0.7609	0.7778	0.9339	1.0000	1.0227	1.1111
RAB	1.0543	0.0761	0.0000	0.7027	1.0000	1.1023	1.5000
ρ^*							
0.00	0.3012	0.0217	0.0000	–0.2008	–0.2857	–0.3149	–0.4286
0.10	0.2711	0.0196	0.0000	–0.1807	–0.2571	–0.2834	–0.3857
0.20	0.2410	0.0174	0.0000	–0.1606	–0.2286	–0.2519	–0.3429
0.30	0.2109	0.0152	0.0000	–0.1405	–0.2000	–0.2205	–0.3000
0.40	0.1807	0.0130	0.0000	–0.1205	–0.1714	–0.1890	–0.2571
0.50	0.1506	0.0109	0.0000	–0.1004	–0.1429	–0.1575	–0.2143
0.60	0.1205	0.0087	0.0000	–0.0803	–0.1143	–0.1260	–0.1714
0.70	0.0904	0.0065	0.0000	–0.0602	–0.0857	–0.0945	–0.1286
0.80	0.0602	0.0043	0.0000	–0.0402	–0.0571	–0.0630	–0.0857
0.90	0.0301	0.0022	0.0000	–0.0201	–0.0286	–0.0315	–0.0429
0.99	0.0030	0.0002	0.0000	–0.0020	–0.0029	–0.0031	–0.0043

RAB = relative absolute bias

biases given in Eqs. A2 (in the Appendix) and 5 that the relative absolute bias $RAB\{\hat{\rho}^*(c)\}$ of estimator ICC(2) with respect to estimator $\hat{\rho}^*(c)$ is the ratio

$$RAB\left\{\hat{\rho}^*(c)\right\} = \frac{\left|Bias\left\{\hat{\rho}^*(c)\right\}\right|}{|Bias\{\mathrm{ICC}(2)\}|} = \frac{|c(N-1)-N+3|}{2}. \quad (6)$$

The MSE formulation in Eq. A4 (Appendix) shows that the relative MSE $RMSE\{\hat{\rho}^*(c)\}$ of estimator ICC(2) with respect to estimator $\hat{\rho}^*(c)$ is the ratio

$$RMSE\left\{\hat{\rho}^*(c)\right\} = \frac{MSE\left\{\hat{\rho}^*(c)\right\}}{MSE\{\mathrm{ICC}(2)\}} = \frac{M_c}{M_1}, \quad (7)$$

where M_c and M_1 are defined in Eqs. A4 and 5, respectively. It is important to note that the two relative indices $RAB\{\hat{\rho}^*(c)\}$ and $RMSE\{\hat{\rho}^*(c)\}$ do not depend on the underlying population ρ^*. For a designated estimator $\hat{\rho}^*(c)$, the associated values of $RAB\{\hat{\rho}^*(c)\}$ and $RMSE\{\hat{\rho}^*(c)\}$ only vary with the selection of N and K, and serve the comparison purpose for the practical situation that the underlying

Table 2 The bias of average score intraclass correlation coefficient indices for $N = 10$ and $K = 50$

$\hat{\rho}^*$	$\hat{\rho}^*_{MS}$	$\hat{\rho}^*_{MO}$	$\hat{\rho}^*_{UB}$	$\hat{\rho}^*_{ME}$	$\hat{\rho}^*_{AV}$	$\hat{\rho}^*_{EF}$	$\hat{\rho}^*_{ML}$
c	0.5533	0.7746	0.7778	0.9283	1.0000	1.0041	1.1111
RAB	1.0102	0.0142	0.0000	0.6771	1.0000	1.0184	1.5000
ρ^*							
0.00	0.2886	0.0041	0.0000	–0.1935	–0.2857	–0.2910	–0.4286
0.10	0.2598	0.0037	0.0000	–0.1741	–0.2571	–0.2619	–0.3857
0.20	0.2309	0.0033	0.0000	–0.1548	–0.2286	–0.2328	–0.3429
0.30	0.2020	0.0028	0.0000	–0.1354	–0.2000	–0.2037	–0.3000
0.40	0.1732	0.0024	0.0000	–0.1161	–0.1714	–0.1746	–0.2571
0.50	0.1443	0.0020	0.0000	–0.0967	–0.1429	–0.1455	–0.2143
0.60	0.1154	0.0016	0.0000	–0.0774	–0.1143	–0.1164	–0.1714
0.70	0.0866	0.0012	0.0000	–0.0580	–0.0857	–0.0873	–0.1286
0.80	0.0577	0.0008	0.0000	–0.0387	–0.0571	–0.0582	–0.0857
0.90	0.0289	0.0004	0.0000	–0.0193	–0.0286	–0.0291	–0.0429
0.99	0.0029	0.0000	0.0000	–0.0019	–0.0029	–0.0029	–0.0043

RAB = relative absolute bias

Table 3 The bias of average score intraclass correlation coefficient indices for $N = 50$ and $K = 10$

$\hat{\rho}^*$	$\hat{\rho}^*_{MS}$	$\hat{\rho}^*_{MO}$	$\hat{\rho}^*_{UB}$	$\hat{\rho}^*_{ME}$	$\hat{\rho}^*_{AV}$	$\hat{\rho}^*_{EF}$	$\hat{\rho}^*_{ML}$
c	0.9143	0.9549	0.9592	0.9879	1.0000	1.0045	1.0204
RAB	1.0996	0.1040	0.0000	0.7034	1.0000	1.1094	1.5000
ρ^*							
0.00	0.0468	0.0044	0.0000	–0.0299	–0.0426	–0.0472	–0.0638
0.10	0.0421	0.0040	0.0000	–0.0269	–0.0383	–0.0425	–0.0574
0.20	0.0374	0.0035	0.0000	–0.0239	–0.0340	–0.0378	–0.0511
0.30	0.0328	0.0031	0.0000	–0.0210	–0.0298	–0.0330	–0.0447
0.40	0.0281	0.0027	0.0000	–0.0180	–0.0255	–0.0283	–0.0383
0.50	0.0234	0.0022	0.0000	–0.0150	–0.0213	–0.0236	–0.0319
0.60	0.0187	0.0018	0.0000	–0.0120	–0.0170	–0.0189	–0.0255
0.70	0.0140	0.0013	0.0000	–0.0090	–0.0128	–0.0142	–0.0191
0.80	0.0094	0.0009	0.0000	–0.0060	–0.0085	–0.0094	–0.0128
0.90	0.0047	0.0004	0.0000	–0.0030	–0.0043	–0.0047	–0.0064
0.99	0.0005	0.0000	0.0000	–0.0003	–0.0004	–0.0005	–0.0006

RAB = relative absolute bias

population ρ^* is unknown. The selected constant *c*, computed relative absolute bias and relative MSE are also reported in the tables.

An inspection of the results in the presented tables shows that the chosen constants of the estimators $\{\hat{\rho}^*_{MS}, \hat{\rho}^*_{MO}, \hat{\rho}^*_{UB}, \hat{\rho}^*_{ME}, \text{ICC}(2), \hat{\rho}^*_{EF}, \hat{\rho}^*_{ML}\}$ have a consistent order that $c_{MS} < c_{MO} < c_{UB} < c_{ME} < c_{AV} < c_{EF} < c_{ML}$ for all four settings of *N* and *K*. Specifically, the actual values of $\{c_{MS}, c_{MO}, c_{UB}, c_{ME}, c_{AV}, c_{EF}, c_{ML}\}$ are

{0.5435, 0.7609, 0.7778, 0.9339, 1.0000, 1.0227, 1.1111} for (*N*, *K*) = (10, 10);

{0.5533, 0.7746, 0.7778, 0.9283, 1.0000, 1.0041, 1.1111} for (*N*, *K*) = (10, 50);

{0.9143, 0.9549, 0.9592, 0.9879, 1.0000, 1.0045, 1.0204} for (*N*, *K*) = (50, 10);

{0.9176, 0.9584, 0.9592, 0.9867, 1.0000, 1.0008, 1.0204} for (*N*, *K*) = (50, 50).

It follows that the resulting values of $\{c_{MS}, c_{MO}, c_{ME}, c_{EF}\}$ vary with (*N*, *K*), the two components $\{c_{UB}, c_{ML}\}$ only depend on *N*, and $c_{AV} = 1$ has a fixed value.

For the relative absolute biases in Tables 1 and 3 with $K = 10$, they reveal the order

Table 4 The bias of average score intraclass correlation coefficient indices for $N = 50$ and $K = 50$

$\hat{\rho}$*	$\hat{\rho}^*_{MS}$	$\hat{\rho}^*_{MO}$	$\hat{\rho}^*_{UB}$	$\hat{\rho}^*_{ME}$	$\hat{\rho}^*_{AV}$	$\hat{\rho}^*_{EF}$	$\hat{\rho}^*_{ML}$
c	0.9176	0.9584	0.9592	0.9867	1.0000	1.0008	1.0204
RAB	1.0184	0.0192	0.0000	0.6741	1.0000	1.0200	1.5000
ρ^*							
0.00	0.0433	0.0008	0.0000	–0.0287	–0.0426	–0.0434	–0.0638
0.10	0.0390	0.0007	0.0000	–0.0258	–0.0383	–0.0391	–0.0574
0.20	0.0347	0.0007	0.0000	–0.0229	–0.0340	–0.0347	–0.0511
0.30	0.0303	0.0006	0.0000	–0.0201	–0.0298	–0.0304	–0.0447
0.40	0.0260	0.0005	0.0000	–0.0172	–0.0255	–0.0260	–0.0383
0.50	0.0217	0.0004	0.0000	–0.0143	–0.0213	–0.0217	–0.0319
0.60	0.0173	0.0003	0.0000	–0.0115	–0.0170	–0.0174	–0.0255
0.70	0.0130	0.0002	0.0000	–0.0086	–0.0128	–0.0130	–0.0191
0.80	0.0087	0.0002	0.0000	–0.0057	–0.0085	–0.0087	–0.0128
0.90	0.0043	0.0001	0.0000	–0.0029	–0.0043	–0.0043	–0.0064
0.99	0.0004	0.0000	0.0000	–0.0003	–0.0004	–0.0004	–0.0006

RAB = relative absolute bias

Table 5 The mean square error of average score intraclass correlation coefficient indices for $N = 10$ and $K = 10$

$\hat{\rho}^*$	$\hat{\rho}^*_{MS}$	$\hat{\rho}^*_{MO}$	$\hat{\rho}^*_{UB}$	$\hat{\rho}^*_{ME}$	$\hat{\rho}^*_{AV}$	$\hat{\rho}^*_{EF}$	$\hat{\rho}^*_{ML}$
c	0.5435	0.7609	0.7778	0.9339	1.0000	1.0227	1.1111
RMSE	0.3793	0.5200	0.5428	0.8333	1.0000	1.0633	1.3389
ρ^*							
0.00	0.3012	0.4130	0.4311	0.6618	0.7942	0.8445	1.0634
0.10	0.2440	0.3345	0.3492	0.5361	0.6433	0.6841	0.8614
0.20	0.1928	0.2643	0.2759	0.4236	0.5083	0.5405	0.6806
0.30	0.1476	0.2024	0.2112	0.3243	0.3892	0.4138	0.5211
0.40	0.1084	0.1487	0.1552	0.2383	0.2859	0.3040	0.3829
0.50	0.0753	0.1033	0.1078	0.1655	0.1986	0.2111	0.2659
0.60	0.0482	0.0661	0.0690	0.1059	0.1271	0.1351	0.1701
0.70	0.0271	0.0372	0.0388	0.0596	0.0715	0.0760	0.0957
0.80	0.0120	0.0165	0.0172	0.0265	0.0318	0.0338	0.0425
0.90	0.0030	0.0041	0.0043	0.0066	0.0079	0.0084	0.0106
0.99	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001

RMSE = relative mean square error

Table 7 The mean square error of average score intraclass correlation coefficient indices for $N = 50$ and $K = 10$

$\hat{\rho}^*$	$\hat{\rho}^*_{MS}$	$\hat{\rho}^*_{MO}$	$\hat{\rho}^*_{UB}$	$\hat{\rho}^*_{ME}$	$\hat{\rho}^*_{AV}$	$\hat{\rho}^*_{EF}$	$\hat{\rho}^*_{ML}$
c	0.9143	0.9549	0.9592	0.9879	1.0000	1.0045	1.0204
RMSE	0.8482	0.8823	0.8898	0.9601	1.0000	1.0162	1.0809
ρ^*							
0.00	0.0468	0.0487	0.0491	0.0530	0.0552	0.0561	0.0596
0.10	0.0379	0.0394	0.0398	0.0429	0.0447	0.0454	0.0483
0.20	0.0299	0.0312	0.0314	0.0339	0.0353	0.0359	0.0382
0.30	0.0229	0.0238	0.0241	0.0260	0.0270	0.0275	0.0292
0.40	0.0168	0.0175	0.0177	0.0191	0.0199	0.0202	0.0215
0.50	0.0117	0.0122	0.0123	0.0132	0.0138	0.0140	0.0149
0.60	0.0075	0.0078	0.0079	0.0085	0.0088	0.0090	0.0095
0.70	0.0042	0.0044	0.0044	0.0048	0.0050	0.0050	0.0054
0.80	0.0019	0.0019	0.0020	0.0021	0.0022	0.0022	0.0024
0.90	0.0005	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006
0.99	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

RMSE = relative mean square error

$RAB\{\hat{\rho}^*_{UB}\} < RAB\{\hat{\rho}^*_{ME}\} < RAB\{\hat{\rho}^*_{MO}\} < RAB\{\text{ICC(2)}\} < RAB\{\hat{\rho}^*_{MS}\} < RAB\{\hat{\rho}^*_{EF}\} < RAB\{\hat{\rho}^*_{ML}\}$.

When $K = 50$, only the relative absolute biases between $\hat{\rho}^*_{ME}$ and $\hat{\rho}^*_{MO}$ are switched in Tables 2 and 4:

$$RAB\{\hat{\rho}^*_{UB}\} < RAB\{\hat{\rho}^*_{MO}\} < RAB\{\hat{\rho}^*_{ME}\} < RAB\{\text{ICC(2)}\} < RAB\{\hat{\rho}^*_{ME}\} < RAB\{\hat{\rho}^*_{EF}\} < RAB\{\hat{\rho}^*_{ML}\}.$$

Obviously, $RAB\{\hat{\rho}^*_{UB}\} = 0$ because $\hat{\rho}^*_{UB}$ is an unbiased index of ρ^*. The resulting biases imply that $\hat{\rho}^*_{MS}$ and $\hat{\rho}^*_{MO}$ are positively biased, while $\hat{\rho}^*_{ME}$, ICC(2), $\hat{\rho}^*_{EF}$, and $\hat{\rho}^*_{ML}$ tend to underestimate ρ^*. In addition, $RAB\{\hat{\rho}^*_{MO}\}$ and $RAB\{\hat{\rho}^*_{ME}\}$ are consistently less than $RAB\{\text{ICC(2)}\} = 1$. It is also interesting to see that $RAB\{\hat{\rho}^*_{MS}\}$ and $RAB\{\hat{\rho}^*_{EF}\}$ are marginally larger than $RAB\{\text{ICC(2)}\}$, and $RAB\{\hat{\rho}^*_{ML}\} = 1.5$ for all N

Table 6 The mean square error of average score intraclass correlation coefficient indices for $N = 10$ and $K = 50$

$\hat{\rho}^*$	$\hat{\rho}^*_{MS}$	$\hat{\rho}^*_{MO}$	$\hat{\rho}^*_{UB}$	$\hat{\rho}^*_{ME}$	$\hat{\rho}^*_{AV}$	$\hat{\rho}^*_{EF}$	$\hat{\rho}^*_{ML}$
c	0.5533	0.7746	0.7778	0.9283	1.0000	1.0041	1.1111
RMSE	0.3837	0.5349	0.5393	0.8179	1.0000	1.0114	1.3448
ρ^*							
0.00	0.2886	0.4024	0.4057	0.6153	0.7522	0.7608	1.0116
0.10	0.2338	0.3259	0.3286	0.4984	0.6093	0.6162	0.8194
0.20	0.1847	0.2575	0.2596	0.3938	0.4814	0.4869	0.6474
0.30	0.1414	0.1972	0.1988	0.3015	0.3686	0.3728	0.4957
0.40	0.1039	0.1449	0.1460	0.2215	0.2708	0.2739	0.3642
0.50	0.0721	0.1006	0.1014	0.1538	0.1881	0.1902	0.2529
0.60	0.0462	0.0644	0.0649	0.0984	0.1204	0.1217	0.1619
0.70	0.0260	0.0362	0.0365	0.0554	0.0677	0.0685	0.0910
0.80	0.0115	0.0161	0.0162	0.0246	0.0301	0.0304	0.0405
0.90	0.0029	0.0040	0.0041	0.0062	0.0075	0.0076	0.0101
0.99	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001

RMSE = relative mean square error

Table 8 The mean square error of average score intraclass correlation coefficient indices for $N = 50$ and $K = 50$

$\hat{\rho}^*$	$\hat{\rho}^*_{MS}$	$\hat{\rho}^*_{MO}$	$\hat{\rho}^*_{UB}$	$\hat{\rho}^*_{ME}$	$\hat{\rho}^*_{AV}$	$\hat{\rho}^*_{EF}$	$\hat{\rho}^*_{ML}$
c	0.9176	0.9584	0.9592	0.9867	1.0000	1.0008	1.0204
RMSE	0.8489	0.8860	0.8874	0.9552	1.0000	1.0030	1.0841
ρ^*							
0.00	0.0433	0.0452	0.0453	0.0488	0.0510	0.0512	0.0553
0.10	0.0351	0.0366	0.0367	0.0395	0.0413	0.0415	0.0448
0.20	0.0277	0.0289	0.0290	0.0312	0.0327	0.0328	0.0354
0.30	0.0212	0.0222	0.0222	0.0239	0.0250	0.0251	0.0271
0.40	0.0156	0.0163	0.0163	0.0176	0.0184	0.0184	0.0199
0.50	0.0108	0.0113	0.0113	0.0122	0.0128	0.0128	0.0138
0.60	0.0069	0.0072	0.0072	0.0078	0.0082	0.0082	0.0089
0.70	0.0039	0.0041	0.0041	0.0044	0.0046	0.0046	0.0050
0.80	0.0017	0.0018	0.0018	0.0020	0.0020	0.0020	0.0022
0.90	0.0004	0.0005	0.0005	0.0005	0.0005	0.0005	0.0006
0.99	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

RMSE = relative mean square error

and K. Hence, $\hat{\rho}^*_{UB}$, $\hat{\rho}^*_{MO}$ and $\hat{\rho}^*_{ME}$ perform better than ICC(2) under the bias principle.

On the other hand, the relative MSEs have the same order between the c constants for all four combinations of N and K:

$$RMSE\{\hat{\rho}^*_{MS}\} < RMSE\{\hat{\rho}^*_{MO}\} < RMSE\{\hat{\rho}^*_{UB}\} < RMSE\{\hat{\rho}^*_{ME}\} < RMSE\{\text{ICC}(2)\} < RMSE\{\hat{\rho}^*_{EF}\} < RMSE\{\hat{\rho}^*_{ML}\}.$$

The consistency indicates that $\hat{\rho}^*_{MS}$ incurs the smallest RMSE, whereas $\hat{\rho}^*_{ML}$ has the largest value. However, there are minor differences between $RMSE\{\hat{\rho}^*_{MO}\}$ and $RMSE\{\hat{\rho}^*_{UB}\}$ or $RMSE\{\text{ICC}(2)\}$ and $RMSE\{\hat{\rho}^*_{EF}\}$. The evidence suggests that ICC(2) is not a prudent selection because $RMSE\{\hat{\rho}^*_{ML}\}$ is the only case that is substantially greater than $RMSE\{\text{ICC}(2)\}$. In short, the four measures $\hat{\rho}^*_{MS}$, $\hat{\rho}^*_{MO}$, $\hat{\rho}^*_{UB}$, and $\hat{\rho}^*_{ME}$ dominate ICC(2) in terms of MSE.

Comparatively, $\hat{\rho}^*_{EF}$ and $\hat{\rho}^*_{ML}$ are the worst two estimators based on both estimation criteria of bias and MSE. Although $\hat{\rho}_{ML}$ has been a competitively accurate index of ρ when the individual rating ICC is small (Shieh, 2012), the counterpart $\hat{\rho}^*_{ML}$ appears to be unsatisfactory in estimating ρ^* regardless of the true magnitude of the average score ICC. More importantly, $\hat{\rho}^*_{MO}$, $\hat{\rho}^*_{UB}$, and $\hat{\rho}^*_{ME}$ outperform ICC(2) under both bias and MSE considerations, and $\hat{\rho}^*_{MS}$ provides a strong alternative to ICC(2) with respect to MSE criterion. The conventional use of ICC(2) for the estimation of mean rating ICC was not supported both analytically and empirically. For the primary reasons of statistical efficiency and computational ease, it is sensible to employ simple and robust alternatives.

In general, the performance of the prescribed estimators improves with an increasing ρ, a larger number of groups N, or a greater group size K when all other features remain constant. The only exception is the case of $\hat{\rho}^*_{UB}$ because it is unbiased for all $N > 3$, $K > 1$, and $0 \le \rho < 1$. Although the specific magnitudes of N and K have a concurrent impact on estimation behavior, the influence of the number of groups differs from that of the group size. Note that the settings (N, K) = (10, 50) and (50, 10) of the one-way random effects model have the identical total sample size 500. But the accuracy and efficiency of the seven estimators in Tables 3 and 7 with (N, K) = (50, 10) are consistently better than the corresponding results in Tables 2 and 6 with (N, K) = (10, 50). Consequently, the discrepancy between the numerical assessments implies that an increase in the number of group, rather than the number of judges in each group, yields more pronounced improvement in estimation for a given total sample size. The particular phenomenon is also confirmed by the additional estimation performance of the two configurations of (N, K) = (20, 25) and (25, 20) with the same total sample size 500. Accordingly, this finding may be useful for researchers to justify their allocation scheme for advance design planning of reliability studies.

Discussion and conclusions

This article concerns the use of ICC(2) as an average score ICC measure within the context of one-way random effects model. Despite its routine and common application in research across many scientific fields, the fundamental properties of the ICC(2) formula is seldom addressed. Although ICC(1) and ICC(2) are conceptually distinct indices, ICC(2) may have been treated as a trivial exercise of the Spearman-Brown equation to ICC(1). This research contributes to the reliability literature by considering the various issues in choosing the best average score ICC index with analytic clarifications and numeric expositions.

First, the estimation of the average score ICC is appropriately recognized as a unique and distinct task in reliability research. The essential attributes of individual rating ICC estimators and the Spearman-Brown formula are synthesized to present a family of average score ICC indices that subsumes ICC(2) as a special case. Accordingly, the choice of the suggested formulation is motivated by its advantages of methodological transparency, analytic tractability, and computational simplicity. Second, exact estimation properties of the suggested class of measures are derived to facilitate the comparison of the strengths and weaknesses of different indices. Within the proposed class of estimators, the best unbiased and the best MSE mean rating ICC indices are identified. Consequently, the theoretical implication and computational ease of the superior alternatives strongly suggest that ICC(2) is sub-optimal and its practical value appears to be difficult to justify.

A potential deficiency of ICC(2) and other average score ICC estimators is that they can assume negative values even though ICC is defined as a non-negative parameter. In practice, the estimate is often set equal to zero when this occurs. Although this simple and intuitive adjustment is of practical meaning, the fundamental behavior of the estimator is inherently altered and a single simple formula of bias and MSE cannot be obtained. However, the conducted Monte Carlo simulation study showed that the bias and MSE performance of the average score ICC indices is essentially unchanged unless the population individual rating ICC is extremely small. A simple explanation is that it is possible to take trivial values of ICC(1) and obtain sizeable values of ICC(2). The occurrence of truncation is less often for the average score ICC indices than the corresponding individual rating ICC estimators. Hence, the numerical details are not reported here.

According to the editorial guidelines and methodological recommendations of several prominent educational and

psychological journals, it is necessary to include some measures of effect size and confidence intervals for all primary outcomes (Alhija & Levy, 2009; Fritz, Morris & Richler, 2012; Kelley & Preacher, 2012; Odgaard & Fowler, 2010; Peng et al., 2013; Sun, Pan & Wang, 2010). Among the several inadequate effect size reporting and interpretation practices, Alhija and Levy (2009) and Peng et al. (2013) especially emphasized that the majority of popular effect size measures are positively biased estimators such as the standardized mean difference index Cohen's d, the strength of association measure $\hat{\eta}^2$, and the sample squared multiple correlation coefficient R^2. These indices are obtained by replacing population parameters with corresponding sample statistics. However, a combination of unbiased component estimators does not necessarily yield an unbiased whole estimator. To expedite the advocated reform of statistical reporting practices, researchers should prudently apply unbiased estimators or other improved formulas in the selection and computation of appropriate effect size measures. Note that unbiasedness is not the only criterion of theoretical importance. Mean square error is another useful performance criterion obtained by incorporating the bias (accuracy) and variability (precision) of an estimator. A thorough explication and comparison of effect sizes under various frameworks certainly facilitate assessment of scientific findings and accumulation of advanced knowledge. This research provides an update and explication of different average score ICC indices that helps to clarify the issue of evaluating the strength of the group property and how to choose an appropriate effect size estimate in multilevel analysis. On the other hand, a thorough coverage of inferential procedures is presented in McGraw and Wong (1996) for hypothesis testing and interval estimation of various average score ICCs.

Acknowledgments The author thanks the action editor, Dale Barr, Pierre Courrieu, and an anonymous reviewer for their helpful comments and suggestions that have improved the presentation of the results in this article.

Appendix

Estimation properties of average score ICC indices

Under the model assumption defined in Eq. 1, the ANOVA F test statistic F^* has the distribution

$$F^* \sim \frac{F\{N-1,\ N(K-1)\}}{1-\rho^*}. \tag{A1}$$

Note that the distribution $F\{N-1, N(K-1)\}$ can be expressed as $[\chi^2(N-1)/(N-1)]/[\chi^2\{N(K-1)\}/\{N(K-1)\}]$ where $\chi^2(\nu)$ is a chi-square distribution with ν degrees of freedom, and the kth moment of $\chi^2(\nu)$ is $E[\{\chi^2(\nu)\}^k] = 2^k\Gamma(\nu/2+k)/\Gamma(\nu/2)$ for $\nu > -2k$. Accordingly, $E[F\{a, b\}] = b/(b-2)$ for $b>2$, and $Var[F\{a, b\}] = \{2b^2(a+b-2)\}/\{a(b-2)^2(b-4)\}$ for $b>4$ as described in Johnson et al. (1995, Chapter 27) and Mood, Graybill, and Boes (1974, Section 4.4).

It follows from the definitions of $\hat{\rho}^*(c)$ and F^* given in Eqs. 4 and A1, respectively, that

$E[\hat{\rho}^*(c)\,] = 1 - c(1-\rho^*)\cdot E[\frac{1}{F\{N-1,\ N(K-1)\}}] = 1 - c(1-\rho^*)\cdot E[F\{N(K-1), N-1\}]$

and

$Var[\hat{\rho}^*(c)\,] = c^2(1-\rho^*)^2\cdot Var[\frac{1}{F\{N-1,\ N(K-1)\}}] = c^2(1-\rho^*)^2\cdot Var[F\{N(K-1), N-1\}]$.

With the prescribed results for the expected value and variance of an F distribution, the expected value $E[\hat{\rho}^*(c)\,]$ and the variance $Var[\hat{\rho}^*(c)\,]$ of an estimator $\hat{\rho}^*(c)$ can be obtained as

$$E\left[\hat{\rho}^*(c)\right] = 1 - \frac{c(N-1)\left(1-\rho^*\right)}{N-3}.$$

and

$$Var\left[\hat{\rho}^*(c)\right] = \frac{2c^2(N-1)^2(NK-3)(1-\rho)^2}{N(N-5)(N-3)^2(K-1)},$$

respectively.

The two basic criteria for comparing the performance of point estimators are based on the bias and mean square error considerations. First, the accuracy of $\hat{\rho}^*(c)$ for the estimation of ρ^* can be verified by the bias $Bias\{\hat{\rho}^*(c)\ \}$ of $\hat{\rho}^*(c)$

$$Bias\left\{\hat{\rho}^*(c)\right\} = E\left[\hat{\rho}^*(c)-\rho\right] = (1-\rho)B_c, \tag{A2}$$

where

$$B_c = 1-\frac{c(N-1)}{N-3}.$$

Thus, it can be readily shown from the bias $Bias\{\hat{\rho}^*(c)\ \}$ that $\hat{\rho}^*_{UB} = \hat{\rho}^*(c_{UB})$ is an unbiased estimator of ρ^* where

$$c_{UB} = \frac{N-3}{N-1}. \tag{A3}$$

Moreover, $\hat{\rho}^*_{UB}$ is a function of (MSB, MSW), and (MSB, MSW) is a complete sufficient statistic for $(K\sigma^2_\gamma+\sigma^2_\varepsilon,\ \sigma^2_\varepsilon)$ as noted in Olkin and Pratt (1958). Consequently, $\hat{\rho}^*_{UB}$ is the best unbiased estimator of ρ^*.

Second, the mean square error $MSE\{\hat{\rho}^*(c)\}$ of $\hat{\rho}^*(c)$ is

$$MSE\left\{\hat{\rho}^*(c)\right\} = E\left[\left(\hat{\rho}^*(c)-\rho^*\right)^2\right] = \left(1-\rho^*\right)^2 M_c, \quad \text{(A4)}$$

where

$$M_c = 1-\frac{2c(N-1)}{N-3}+\frac{c^2(N-1)^2\{N(K-1)+2\}}{N(N-3)(N-5)(K-1)}.$$

It is straightforward to show that M_c attains the minimum for $c = c_{MS}$ where

$$c_{MS} = \frac{N(N-5)(K-1)}{(N-1)\{N(K-1)+2\}}. \quad \text{(A5)}$$

Therefore, $\hat{\rho}^*_{MS} = \hat{\rho}^*(c_{MS})$ is the best MSE index within the class of estimators $\hat{\rho}^*(c)$. Apparently, the prescribed estimation results are valid for the values of $N > 5$ and $K > 1$ that are likely to be satisfied in actual applications.

References

Alhija, F. N. A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement, 69,* 245–265.

Bartko, J. J. (1976). On various intraclass correlation reliability. *Psychological Bulletin, 83,* 762–765.

Bliese, P. D. (1998). Group size, ICC values and group-level correlations: A simulation. *Organizational Research Methods, 1,* 355–373.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco: Jossey-Bass.

Bliese, P. D., & Halverson, R. R. (1998). Group size and measurements of group-level properties: An examination of eta-squared and ICC values. *Journal of Management, 24,* 157–172.

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

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review, 54,* 67–82.

Donoghue, J. R., & Collins, L. M. (1990). A note on the unbiased estimation of the intraclass correlation. *Psychometrika, 55,* 159–164.

Fisher, R. A. (1938). *Statistical Methods for Research Workers* (7th ed.). Edinburgh: Oliver and Boyd.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141,* 2–18.

Gleason, J. R. (1997). sg65: Computing intraclass correlations and large ANOVAs. *Stata Technical Bulletin, 35,* 25–31.

Harris, I. R., & Burch, B. D. (2000). Pivotal estimation with applications for the intraclass correlation coefficient in the balanced one-way random effects model. *Journal of Statistical Planning and Inference, 83,* 257–276.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29,* 60–87.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Psychological Bulletin, 67,* 219–239.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York: Wiley.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137–152.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9,* 202–220.

LeBrenton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K. P., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6,* 80–128.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815–852.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

Odgaard, E. C., & Fowler, R. L. (2010). Confidence intervals for effect sizes: Compliance and clinical significance in the *Journal of Consulting and Clinical Psychology. Journal of Consulting and Clinical Psychology, 78,* 287–297.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201–211.

Peng, C. Y. J., Chen, L. T., Chiang, H. M., & Chiang, Y. C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychological Review, 25,* 157–209.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*. New York: Wiley.

Shieh, G. (2012). A comparison of two indices for the intraclass correlation coefficient. *Behavior Research Methods, 44,* 1212–1223.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 171–195.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102,* 989–1004.