Choice of Agreement Statistics: A Discussion of the Underlying Biostatistics, With Heuristic Examples From the Vineland™–3

Professor Dom Cicchetti, PhD
Department of Biometry
Yale University School of Medicine
New Haven, CT 06520
Fellow, American Statistical Association

Abstract

This report discusses

- the difference between correlation and agreement;
- the rationale underlying the application of an appropriate model of the intra-class correlation coefficient;
- the arguments underlying the application of Cohen's *d* and *r* statistics as measures of effect sizes (ES), in the context of defining levels of clinical or practical significance, as compared to statistical significance (e.g., Borenstein, 1998; Cicchetti, 2008; Cohen, 1988); and
- DOMENIC, a novel statistic developed by the author that can be utilized to help resolve challenging classification problems.

Choice of Agreement Statistics: A Discussion of the Underlying Biostatistics, With Heuristic Examples From the Vineland™–3

Distinguishing Between Correlation and Agreement

Famed statistician Sir Ronald A. Fisher made the distinction between correlation and agreement in his classic text, *Statistical Methods for Research Workers* in 1938, yet the application of the incorrect correlational model persists. Using the standard Pearson product–moment correlation coefficient (PPMC) to demonstrate agreement rather than an appropriate model of the intra-class correlation coefficient (ICC) is a common problem. While Vineland users can rest assured that inter-rater agreement levels are high, the author has experienced this problem in his work as a biostatistical consultant to various journals across the disciplines of psychology, neuropsychology, medicine, psychiatry, and oenology, as well as in his current role as statistical editor for the *Journal of Nervous and Mental Disease.*

One can legitimately ask, why should the problem still exist? There seem to be two reasons why this incorrect strategy continues to be employed. First, the PPMC is readily available and much easier to understand. A second, but much more subtle, reason is that some research scientists have applied the PPMC and an appropriate model of the ICC to the same data set and have obtained very similar results. When there is, in fact, a high level of agreement between any given pair of raters, the PPMC and an appropriate model of the ICC will indeed produce similar results. The point here is that the PPMC simply measures the extent to which pairs of raters' scores vary in the same order, *not the extent to which the raters' individual scores actually disagree with each other* (Kazdin, 1982).

The PPMC is appropriate for observing the association between two variables when each one is measured on a *different* scale, such as the correlation between height in inches and weight in pounds or the correlation between the number of gallons of gas in an automobile and the number of miles a driver can travel. Conversely, when two or more examiners apply the same scale of measurement, the interest is in the extent of chance-corrected rater agreement, NOT in the correlation or covariation between the variables of interest.

To highlight the differences between the PPMC and the ICC, consider the following heuristic examples for three different conditions:

- when the ICC results in a higher value than the PPMC;
- when the ICC and the PPMC produce similar results and, finally,
- when the PPMC produces consistently higher values than the ICC.

The data in the examples were simulated to lie within ± 1 standard deviation of what would be considered a normal range of Vineland standard scores, here between 86 and 111. The ICC coefficients were calculated using Model (2,1) (Shrout & Fleiss, 1979). Results in these examples were interpreted according to the following guidelines (Cicchetti, 1994; Cicchetti & Sparrow, 1981):
< 0.40 = Poor (P); 0.40–0.59 = Fair (F); 0.60–0.74 = Good (G); and 0.75–1.00 = Excellent (E).

When the ICC Results are Higher Than the PPMC

In this example, the ICC value of 0.40 (Fair) is greater than the PPMC value of 0.35 (Poor).

Clinician 1	Clinician 2
100	101
102	106
107	111
101	100
106	107
111	102

When the ICC and the PPMC Produce Similar Results

The ICC and PPMC values are virtually the same, with the ICC producing a coefficient of 0.60 and the PPMC a value of 0.61 (Good).

Clinician 1	Clinician 2
105	105
104	105
104	104
104	104
104	104

When the PPMC Produces Consistently Higher Values Than the ICC

Though it is possible for the ICC to have higher values than the PPMC, the more serious problem occurs when the PPMC is consistently *much* higher than the ICC. According to Kazdin's (1982) caveat, the size of the correlation coefficient depends upon the extent to which two sets of scores co-vary, irrespective of how far apart the pairings of the scores happen to be. For PPMC purposes, this means that when five scores (e.g., 50, 55, 60, 65, and 70) are paired with the same five values in the same order, the result is a perfect positive correlation of +1.00; however, when these same five values are paired with a set of very different values (e.g., 1, 2, 3, 4, and 5), the same PPMC coefficient of +1.00 is produced. The ICC (2,1) model, on the other hand, produces a coefficient of 0.01, which makes biostatistical, as well as clinical, sense, given how far apart the paired values happen to be.

The PPMC Ignores the Extent of Disagreement Between Pairs of Evaluators

To illustrate this phenomenon more precisely, consider the data presented in Table 1.

Table 1. Hypothetical Vineland™–3 Domain Data for Comparison of the PPMC and the ICC

A	A′	B	C	D	E	F	G	H	I	J	K
105	105	104	103	102	101	100	99	98	97	96	95
104	104	103	102	101	100	99	98	97	96	95	94
103	103	102	101	100	99	98	97	96	95	94	93
102	102	101	100	99	98	97	96	95	94	93	92
101	101	100	99	98	97	96	95	94	93	92	91
100	100	99	98	97	96	95	94	93	92	91	90
99	99	98	97	96	95	94	93	92	91	90	89
98	98	97	96	95	94	93	92	91	90	89	88
97	97	96	95	94	93	92	91	90	89	88	87
96	96	95	94	93	92	91	90	89	88	87	86

Note. For each possible pairing with A, namely, AB, AC, AD, AE, AF AG, AH, AI, AJ and AK, the PPMC produces a perfect +1.00 value; however, when the appropriate model of the intra-class correlation coefficient (ICC) is applied, the only pairing that reaches a value of +1.00 is between A and A′. As the level of agreement decreases systematically from B to C to D to E to F to G, to H, to I, to J, to K, so does the value of the ICC.

In pairing Case A with each of the remaining cases, namely, A′, B, C, D, E, F, G, H, I, J, and K, each PPMC coefficient is +1.00 (or a perfect positive correlation); on the other hand, the ICC (2,1) becomes progressively lower as the level of agreement decreases, relative to Case A. Specifically: AB = 0.95 (E); AC = 0.82 (E); AD = 0.67 (G); AE = 0.53 (F); AF = 0.42 (F); AG = 0.34 (P); AH = 0.27 (P); AI = 0.22 (P); AJ = 0.18 (P); and AK = 0.15 (P). These results demonstrate the perils of reporting PPMC correlations instead of the ICC correlation coefficients.

At this point in the argument, it is logical to ask, how does the application of the PPMC, rather than the ICC (2,1), occur? The answer appears to be that the responsible clinical research scientist ensures, in any given inter-examiner reliability study, that the assessors/examiners/raters/clinicians are adequately informed and trained to make appropriate assessments. When this occurs, it increases the probability that agreement levels are high; and as demonstrated by the second example, PPMC and ICC (2,1) will produce similar results. To demonstrate this phenomenon, the inter-examiner reliability of the Vineland–3 was assessed with the ICC (2,1) and the PPMC. Because the raw data showed a high level of chance-corrected inter-examiner agreement, the ICC (2,1) and the PPMC produced nearly identical results.

All this said, it is a serious mistake to assume that whenever two methods produce the same results the methods are equivalent. This is false and is an example of what might be referred to as pseudo-equivalence. *The conceptual problem here is that it is not possible, a priori, to be sure that the examiners have been adequately trained; and it is therefore not possible to know whether inter-rater agreement will be high and acceptable before the results have occurred.*

It is important to understand that these arguments should not be interpreted as a denigration of the venerable PPMC, a valuable statistic that has survived the test of time and, in fact, has been utilized in a variety of creative contexts as noted by well-regarded biostatisticians. Rodgers and Nicewander (1988) identified the following 13 ways that the PPMC has been interpreted:

1. As Pearson (1896) defined it or as it is typically applied
2. As a ratio of standard deviations
3. As the standardized slope of the regression line
4. As the geometric average of the two regression slopes
5. As the proportion of variance accounted for
6. As the average cross product of standardized variables
7. In relation to the angle between two standard regression lines
8. In relation to the angle between the two variable vectors
9. As a rescaled variance of the difference between standardized scores
10. As estimated from the balloon rule: Note that the balloon is formed by drawing an ellipse around the scatterplot of the individual X and Y values
11. As a more formal representation of the balloon rule
12. As related to test statistics from designed experiments
13. As the ratio of two means

Rovine and von Eye (1997) added a 14th showing the PPMC has been interpreted as:

14. The proportion of matches between standardized X and Y values

Using *d* as a Measure of Effect Size (ES)

Cohen's *d* is widely used to determine the level of clinical significance of measures of intra- and inter-examiner agreement, as well as correlation or association more generally. It can be defined as the difference between two averages or mean scores divided by the standard deviation *(SD)* of either one of the two means, because they are assumed to be equal (Cohen, 1988, p. 20):

$$d = \frac{M_A - M_B}{SD} \qquad (1)$$

As in previous editions of the Vineland, the Vineland–3 reported effect sizes using the *d* statistic. The author's preference is to use *r* because of its familiarity to the non-biostatistician. However, the solution is simple; *r* is easily obtained by transforming the *d* statistic as shown below (Cohen, 1988, p. 23); *d* is defined in Equation 1.

$$r = \frac{d}{\sqrt{(d^2 + 4)}} \qquad (2)$$

In deciding between *d* and *r,* McGrath and Meyer (2006, p. 398) offered this solution:

> A final option is to report *d* as well as *r.* Doing so has several benefits, including simplicity and the fact that it does not require adjusting interpretive benchmarks. An additional benefit is that when base rates diverge, reporting both *r* and *d* will juxtapose the seemingly discrepant inferences about magnitude of effect and will highlight the importance of deciding whether the natural base rates should be given credence or be discounted. However, for efficiency, researchers may prefer adjusting the base rates in instances in which large numbers of effect-size statistics are reported for a single sample.

To aid interpreting *d* and *r,* Cohen (1988) developed two sets of criteria for effect size interpretation. The *r criteria* were later expanded by Cicchetti in 2008. The data presented in Table 2, which are from both Cohen (1988) and Zakzanis (2001), shows the relationship between *d* and *r*. Cohen's criteria and the author's expansion are presented in the table note.

Table 2. Relationship between Cohen's *d* and *r*

d[a]	% Overlap	*r*[a]
0.0	100.0	0.000
0.1	92.3	0.050
0.2	85.3	0.100
0.3	78.7	0.148
0.4	72.6	0.196
0.5	66.6	0.243
0.6	61.8	0.287
0.7	57.0	0.330
0.8	52.6	0.371
0.9	48.4	0.410
1.0	44.6	0.447
1.1	41.1	0.482
1.2	37.8	0.514
1.3	34.7	0.545
1.4	31.9	0.573
1.5	29.3	0.600
1.6	26.9	0.625
1.7	24.6	0.648
1.8	22.6	0.669
1.9	20.6	0.689
2.0	18.9	0.707
2.2	15.7	0.740
2.4	13.0	0.768
2.6	10.7	0.793
2.8	8.8	0.814
3.0	7.2	0.832
3.2	5.8	0.848
3.4	4.7	0.862
3.6	3.7	0.874
3.8	3.0	0.885
4.0	2.3	0.894

[a]The Cohen (1988) criteria for *d* are: < 0.2 = No Effect; 0.2 = Small; 0.5 = Medium; and ≥0.8 = Large. The Cohen (1988) criteria for *r* are: < 0.10 = Trivial; 0.10–0.29 = Small; 0.30–0.49 = Medium; and ≥0.50 = Large. These were revised as: < 0.10 = Trivial; 0.10–0.29 = Small; 0.30–0.49 = Medium; 0.50–0.69 = Large; and ≥ 0.70 = Very Large (Cicchetti, 2008).

Assessing the Adaptive Level of a Person Who Presents a Diagnostic or Classification Challenge: A Hypothetical Example Applying the DOMENIC Reliability Statistic

For this hypothetical case, a child has one or more clinical disorders that negatively affect his or her social functioning, making it difficult to evaluate his or her level of adaptive behavior. The interfering problems may be ADHD, autism spectrum disorder, or some combination of both. The child may demonstrate different levels of adaptive behavior at different hours of the day, depending upon the perceived personalities of the other persons with whom he or she interacts. A potential plan for evaluating and understanding this child may be to select a group of, say, six experts who have experience with the child in a variety of settings, and then interview each of them as to the overall or typical behavior of the child across the diverse social settings.

In this example, the five adaptive levels used in Vineland–II are:

High (H): 2 or more *SD*s above
Moderately High (MH): 1.0 – < 2.0
Adequate (A): –1.0 – < 1.0
Moderately Low (ML): –2.0 – < –1.0
Low (L): below –2

With six examiners, and designating *k* to represent them, the number of pairs of inter-examiner comparisons is given by the formula k(k–1)/2; here equaling (6 x 5)/2 = 30/2 = 15. Further assume that the results are distributed as follows, with the linear weights derived from Cicchetti's 5-category ordinal scale (1976); and where H = High; MH = Moderately High; A = Adequate; ML = Moderately Low; and L = Low.

Table 3. Inter-Examiner Comparisons Based on Cicchetti's 5-Category Ordinal Scale

	Complete agreement 1.00	**One category apart .75**	**Two categories apart .50**	**Three categories apart .25**	**Complete disagreement 0.0**
	H–H = 2 MH–MH = 6 L–L = 3	L–ML = 3 H–MH = 1			
SUMS:	11	4			

With 11 judgments in complete agreement and four of them one category apart, the level of inter-examiner agreement becomes [(11 x 1) + (0.75 X 4)] = 14/15 = 93.33; this represents Excellent agreement, according to the Cicchetti, Volkmar, Klin, and Showalter (1995) criteria, whereby < 70 = Poor (P); 70–79 = Fair (F); 80–89 = Good (G); and > 90 = Excellent (E) agreement. This statistical approach is known as DOMENIC for: the **D**etection **O**f **M**ultiple **E**xaminers **N**ot **I**n **C**onsensus (Cicchetti, 2006).

The level of statistical significance of inter-examiner agreement is calculated by comparing the average agreement level (here 93.33) to 70% (the lowest level of acceptable agreement, based on Cicchetti, Volkmar, Klin, and Showalter criteria, 1995). The second step is to divide this difference by the *SEM* of the k (k–1)/2 pairings:

$$z = (\text{Mean} - 0.70) / SEM \qquad (3)$$

The standard error of the mean *(SEM)* is the *SD* of the agreement weights of each examiner pairing, divided by the square root of the 15 pairings *(N).* The two-tailed *z* score is then evaluated for the level of statistical significance *(p)* as follows:

z	*p*
± 1.645	0.10
± 1.960	0.05
± 2.575	0.01
± 2.960	0.003
± 4.000	< 0.0005
≥ ± 5.000	< 0.0001

The *SEM* for the 15 pairings is 0.0295. Therefore, $z = (.9333 - 0.70)/0.0295 = 7.91$. Because 7.91 is > 5, $p = < 0.0001$. Therefore, in this hypothetical example, 93% agreement is both statistically and clinically significant. For recent research showing the correlation between the reliability and validity of human judgments, see Cicchetti (2011, 2017). While both publications pertain to the reliability and accuracy of diagnoses of autism, the latter publication also provides data to indicate why Cohen's 1960 Kappa statistics should be the statistic of choice for binary diagnostic assessments.

The research of Cicchetti, Showalter, and Tyrer (1985) showed that scales of seven categories or more can be treated as interval scales. This prompted the following statement from NIH:

> New scoring procedures for evaluating research applications for potential FY 2010 grant funding will be based upon the research findings of Cicchetti, Showalter, and Tyrer (1985).

As recently as 2015, the new scoring procedures were still being utilized by NIH.

Conclusions

In this discussion of the choice of agreement statistics, there are four broad areas of discussion. First, the conceptual framework underlying the application of an appropriate model of the intra-class correlation coefficient (ICC); how it compares with the familiar and standard Pearson product–moment correlation coefficient (PPMC); why PPMC is often confused with the ICC; and specific examples illustrating why PPMC is an invalid measure of inter-examiner or intra-examiner agreement are examined. Second, a comparison of *d* and *r* as measures of clinical significance or effect size (ES) is given. Third, the DOMENIC reliability statistic is presented, including a hypothetical example of a diagnostic classification challenge. Fourth, the biostatistical relationship between the reliability and validity of human judgments is cited and referenced. This article would not be complete without acknowledging the outstanding differentiation between statistical and clinical significance offered by Borenstein (1998) as a tribute to the late and great Jacob Cohen.

References

Borenstein, M. (1998). The shift from significance testing to effect size estimation. In A. S. Bellak & M. Hersen (Eds.), *Comprehensive clinical psychology* (vol. *3*, pp. 313–349). New York, NY: Pergamon.

Cicchetti, D.V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry, 129,* 452–456.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Cicchetti, D. V. (2006). The Paris 1976 wine tastings revisited once more. *Journal of Wine Economics, 1,* 125–140.

Cicchetti, D. V. (2008). From Bayes to the just noticeable difference to effect sizes: A note to understanding the clinical and statistical significance of oenologic research findings. *Journal of Wine Economics, 3,* 185–193.

Cicchetti, D. V. (2011). On the reliability and accuracy of the evaluative method for identifying evidence-based practices in autism. In B. Reichow, P. Doehring, D. V. Cicchetti, & F. R. Volkmar (Eds.), *Evidence-based practices and treatments for children with autism* (pp. 41–51). New York, NY: Springer.

Cicchetti, D. V. (2017). Evaluating the value of replicate tastings of a given wine: Bio-statistical considerations. *Journal of Wine Research, 28*(2), 135–143. doi:10.1080/09571264.2017.1312317

Cicchetti, D. V., Showalter, D., & Tyrer, P. (1985). The effect of number of rating scale categories upon levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement, 9,* 31–36.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86,* 127–137.

Cicchetti, D. V., Volkmar, F., Klin, A., & Showalter, D. (1995). Diagnosing autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. *Child Neuropsychology, 1,* 26–37.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Fisher, R. A. (1938). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver & Boyd.

Kazdin, A. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11,* 386–401. doi:10.1037/1082-989X.11.4.386

Rodgers, J. L., & Nicewander, A. (1982). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42,* 59–66.

Rovine, M., & von Eye, A. (1997). A 14th way to look at a correlation coefficient: Correlation as the proportion of matches. *The American Statistician, 51,* 42–46.

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

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland II: A Revision of the Vineland Adaptive Behavior Scales I* Survey/Caregiver Form (2nd ed.). Circle Pines, MN: American Guidance Service.

Zakzanis, K. K. (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: Formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology, 16,* 653–667.