Establishing Reliability When Multiple Examiners Evaluate a Single Case-Part II: Applications to Symptoms of Post-Traumatic Stress Disorder (PTSD)

Domenic Cicchetti[1,*], Alan Fontana[2,3] and Donald Showalter[2]

[1]*Child Study Center and Departments of Biometry and Psychiatry, Yale University School of Medicine, USA*

[2]*North East Program Evaluation Center (NEPEC), West Haven, CT, USA*

[3]*Department of Psychiatry, Yale University School of Medicine, USA*

Abstract: In an earlier article, the authors assessed the clinical significance of each of 19 Clinician Administered PTSD Scale items and composite scores (CAPS-1) [1] when 12 clinicians evaluated a Vietnam era veteran. A second patient was also evaluated by the same 12 clinicians and used for cross-validation purposes [2]. The objectives of this follow-up research are: (1) to describe and apply novel bio-statistical methods for establishing the statistical significance of these reliability estimates when the same 12 examiners evaluated each of the two Vietnam era patients. This approach is also utilized within the broader contexts of the ideographic and nomothetic conceptualizations to science, and the interplay between statistical and clinical or practical significance; (2) to detail the steps for applying the new methodology; and (3) to investigate whether the quality of the symptoms (frequency, intensity); item content; or specific clinician affect the levels of rater reliability. The more typical (nomothetic) reliability research design focuses on group averages and broader principles related to biomedical issues, rather than the focus on the individual case (ideographic approach). Both research designs (ideographic and nomothetic) have been incorporated in this follow-up research endeavor.

Keywords: Multiple Raters, Single Case, Nomothetic, Ideographic, PTSD, Statistical Significance, Clinical Significance.

INTRODUCTION

In an earlier article, the authors assessed the clinical significance of each of 19 Clinician Administered PTSD Scale (CAPS-1) items and composite scores when 12 clinicians evaluated a Vietnam era veteran. A second patient was also evaluated by the same 12 clinicians and used for cross-validation purposes [2]. The objective of this follow-up research is to describe and apply novel bio-statistical methods for establishing the statistical significance of these reliability estimates when the same 12 examiners evaluated each of the two Vietnam era patients. This endeavor will be pursued within the broader context of theoretical issues and applications of this new methodology. Detailed steps will be taken to illustrate precisely how the new methodology can be applied within the broader framework of the critical interplay between clinical and statistical significance.

Theoretical Issues

The ideographic-nomothetic distinction was introduced by the philosopher Windelband [3,4] and popularized in the field of psychology by Gordon Allport, (1937) [5]. From a theoretical perspective, the approach taken in this research is representative of the ideographic model with its focus on the individual. The more prototypic research focus is based upon group averages. By this basic definition, the doctor-patient relationship would represent an ideographic focus upon the individual, per se. The clinician researcher would, in general, be more interested in group phenomena, laws of bio-behavioral functioning, or a so-called nomothetic focus. However, in carrying the argument further, it becomes clear that both approaches are necessary to do meaningful clinical research in whatever biomedical field of inquiry [6]; and, more recently, Robinson (2011) [7].

A prototypical biomedical example of this phenomenon occurs in the evaluation of the health of a patient. The physician considers the patient *both* as a unique individual (ideographic reasoning); but also as someone whose health is measured by state-of-the-art criteria that continue to evolve, as derived from a larger sampling of representative persons in the general population (nomothetic reasoning).

At a more general methodological level, the acclaimed statistician, Sir Ronald A. Fisher, showed the value of the ideographic model, in his carefully designed study of the “lady tasting tea.” To summarize briefly, Fisher tested the ability of a lady to be able to successfully distinguish between cups of tea in which the tea was poured before the milk from those in which the milk was poured first. As his biographer [8] noted:

*Address correspondence to this author at the Child Study Center and Departments of Biometry and Psychiatry, Yale University School of Medicine, USA; Tel: 203 488-6563; Fax: 203 488-4218; E-mail: dom.cicchetti@yale.edu

E-ISSN: 1929-6029/15

"Fisher explains the notions of adequate design, tests of significance, randomization, and sensitivity, all through the lady tasting tea example (p. 36)."

But, whether the data derive from ideographic or nomothetic types of research designs, or both, the clinician/researcher must decide whether the obtained results are statistically meaningful and, beyond this, whether the results have any clinical or practical utility.

Distinguishing Between Statistical and Clinical Significance

The concept of statistical significance is based upon a probability model. Put simply, statistical significance answers the question: what is the probability that a clinical phenomenon/research result occurred on the basis of chance alone? The standard most often adopted by the research scientist is that a given result is statistically significant if it occurred with a chance probability (*p*) of 5% or less.

With respect to clinical significance, the issue being addressed is whether a given scientific result has any clinical or practical significance above and beyond its level of statistical significance.

The Interplay between Statistical and Clinical Significance

There are four possibilities here, each with its own level of substantive meaning. Thus, a research scientist is faced with one of these results: statistically and clinically non-significant; statistically significant, but not clinically significant; clinically significant, but not statistically significant; and both statistically and clinically significant.

While three of the four possible combinations between statistical and clinical significance are easy to understand within the conceptual framework of designing a given research study, the remaining one is not always identified correctly. It is the situation in which a result that appears statistically meaningless at first examination becomes clinically relevant when examined appropriately. An exemplar of this phenomenon derives from cardiology research. Kelly and Preacher (2012, p. 139) [9] recall the results of an earlier research study investigating the effect of aspirin intake (yes/no) upon the occurrence of myocardial infarction (yes/no). The correlation between these two variables was only 0.033, which, by the effect size (ES) criteria of [10] would be considered Trivial, since it is well below 0.10. However, when this ES is interpreted as an odds-ratio, the group receiving no aspirin was almost twice as likely to suffer a heart attack –odds ratio= 1.82- than the group receiving aspirin. Thus, while the desideratum remains that a research result be both statistically and clinically significant, it can also be true that results that seem trivial can demonstrate high clinical relevance when an appropriate statistical test is employed.

When a research finding is clinically significant, but not statistically significant, this often creates a problem, because its meaning is often misinterpreted. As shown in the seminal paper by Borenstein (1998) [11], researchers often confuse statistical with clinical significance. Consider the case in which six drugs or medications have been studied in experimental and control groups to treat a particular disease. Each of them elicits a positive response, favoring the treatment group, at approximately the same acceptable level of clinical significance. Now suppose the three studies with the smallest sample sizes produce results that fail to reach statistical significance at the *p* level of 0.05 or less; while the remaining three studies produce results that are statistically significant. The typical response is to remove the drug from the market with the reasoning, albeit incorrect, that the evidence proves that the drug is not effective. As Borenstein (1998) [11] correctly concludes, the relevant studies should be replicated with larger sample sizes. If the results are now both statistically and clinically significant, this should be interpreted as evidence that the drug is indeed clinically useful. This example also underscores the importance of comparable sample sizes when interpreting levels of statistical and clinical significance.

Another error that has been articulated by Borenstein is unfortunately common among both students and researchers. This is the phenomenon of confusing the size of the *p* value with its level of clinical significance. Thus, a research result with a *p* value of 0.0005 is viewed incorrectly, as being more clinically meaningful than one with a *p* value of 0.05. As with the earlier example, the difference in *p* values is often the result of a difference in sample sizes. While a p value of 0.0005 is more striking than one at the 0.05 level neither should be confused with clinical significance.

The next section will focus upon the general formulae for assessing levels of inter-rater agreement, whether the variables are measured on nominal or ordinal scales; and when multiple raters independently evaluate a single case.

METHODS

Establishing Levels of Inter-Rater Reliability

As introduced by Cohen (1960; 1968) [12, 13], kappa and weighted kappa, respectively, are comprised of three fundamental components: the Proportion/Percentage of Observed inter-rater agreement (PO); the Proportion or Percentage of inter-rater agreement expected on the basis of Chance alone (PC); and the level of chance-corrected agreement (PO-PC)/ (1-PC). This defines Kappa (K) (for nominal variables), or weighted Kappa (K_w) (for ordinal variables), as follows:

$$\text{Kappa (K) /Weighted Kappa } (K_w) = (PO\text{-}PC)/(1\text{-}PC) \quad (1)$$

The kappa statistic (weighted or un-weighted) assumes: a positive value when PO exceeds PC, producing a value of +1.00 when there is perfect or complete inter-rater agreement; a value of 0 when observed and chance agreement are identical (PO-PC)=0; and a negative value when PO is less than PC. In this instance, (PO-PC) = a negative value that can be -1.00 or lower.

In the next section, the authors will discuss various conceptually related sets of guidelines for determining the extent to which levels of kappa or weighted kappa are clinically significant.

Determining Levels of Inter-Rater Reliability that are Clinically Significant

The first set of suggested criteria for determining levels of clinical significance, or the connotatively similar phrase "strength of agreement" were published by Landis & Koch (1977) [14]. The six levels, based upon chance-corrected agreement, were: <0.00=Poor; 0.00-0.20=Slight; 0.21-0.40=Fair; 0.41-0.60=Moderate; 0.61-0.80=Substantial; and 0.81-1.00= an Almost Perfect to a Perfect level of chance-corrected inter-rater agreement.

Four years later, [15] suggested conceptually similar criteria, albeit with fewer levels of inter-rater agreement than were recommended by Landis and Koch, as shown here: < 0.40=Poor; 0.40-0.75=Fair to Moderate; and >0.75=excellent agreement beyond chance. These three levels were suggested again by Fleiss *et al*., (2003, p.605) [16].

Cicchetti & Sparrow (1981) [17] published guidelines that are conceptually similar to those of Fleiss and colleagues, the main difference being that 0.40-0.74 was divided into two categories, resulting in <0.40=Poor; 0.40-0.59= Fair; 0.60-0.74= Good; and ≥ 0.75=Excellent.

Note the conceptually similar relationships between the Cicchetti & Sparrow and Fleiss and colleagues' criteria, on the one hand, and the Landis & Koch criteria, on the other, especially at the higher end, which would be most relevant for defining acceptable levels of chance corrected inter-rater reliability: First, Landis & Koch's Moderate, at 0.41-0.60 maps very closely onto Cicchetti & Sparrow's Fair agreement that ranges between 0.40 and 0.59. Second, Cicchetti & Sparrow's Good agreement, set between 0.60 and 0.74, agrees well with Landis & Koch's Substantial, at 0.61-0.80; and third, Cicchetti & Sparrow's Excellent, at 0.75-1.00 is conceptually similar to the Landis & Koch Almost Perfect to Perfect agreement category ranging between 0.81 and 1.00.

As noted in our previous paper [2], the value of 0.70 has been selected for PC (the Proportion/percentage of Chance agreement). The chance-corrected level of agreement (Kappa for Nominal variables, Weighted Kappa for ordinal variables) of 0.40 is considered to be of clinical significance (Fair/Average) by the criteria of Cicchetti & Sparrow (1981) [17]; Fleiss *et al.,* 2003 [16]; and the earlier criteria of Landis & Koch (1977) [14]. The level of 0.40 will result when PO (the Proportion of Observed) rater agreement) is 0.82, which is considered good agreement by the criteria of Cicchetti, Volkmar, Klin, & Showalter (1995) [18].

In addition, if one considers all the possible levels of agreement that can occur between 10 clinical examiners, diagnosing the presence or absence of PTSD symptomatology (or any other disorder or disease); their average level of inter-rater agreement becomes 0.70 Cicchetti *et al*., (2006, p.563) [19]. It should also be noted that the 70% criterion is an integral part of our school or academic experience wherein it is usually considered the minimal passing grade. Capitalizing on this idea, Robert Parker and some of his enological colleagues use a rating of 70% as the lowest score describing a wine of acceptable quality.

A second and more standard approach would be to base PC on the squaring of the proportion of rater pairings at each level of inter-rater agreement, multiplying the individual rater pairings by its appropriate weight, summing the resulting products

and dividing by (1-PC). For example, PC for rating the *Intensity* of Falling/Staying asleep produced a PC value of 0.40, meaning a level of chance probability level of far less than the results of a coin-flip. When the PO of 0.90 is compared to PC, K_w becomes (.90-.40)/.60= 0.83 which is Excellent by the criteria of Cicchetti & Sparrow (1981) [17] and Almost Perfect, by the earlier criteria of Landis & Koch (1977) [14]. In contrast, when we use .70 as the criterion, K_w becomes (.90-.70)/.30= .67, or Good by Cicchetti & Sparrow, and Substantial according to Landis & Koch. While the reliability researcher is obviously free to use either of these two options, we would, for all the arguments presented here opt for using the .70 criterion for PC.

Finally, Szalai (1993; 1998) [20, 21] to our knowledge, is the only author, other than the current ones who has focused upon the reliability problem when multiple raters evaluate a single case. The author refers to a calculation of PC on the basis of an "equiprobable" model, as a "reasonable" one. For our PTSD variables, this would mean that we would expect, on the basis of chance alone, that the 12 examiners would be as likely to be in complete agreement as they would to be at any level of partial agreement or complete disagreement. Given the high level of training each rater receives in any meaningful reliability research enterprise, this would seem highly unlikely. For this reason, the Szalai calculation of PC is not recommended.

Determining Levels of Inter-Rater Reliability that are Statistically Significant

As creatively conceptualized by Borenstein (1998, p. 315) [11], while there is variation in the specific formulae for testing levels of statistical significance, from one procedure to another, they are, nonetheless, "variations on the theme:

$$\text{Test statistic} = \frac{\text{Observed difference}}{\text{Dispersion of the difference}} \quad (2)$$

In the specific area of deciding whether there is a statistically significant level of agreement when multiple judges evaluate a single case, the specific formula becomes:

$$t/Z^{*} = (PO - PC)/\ SEM, \text{ whereby} \quad (3)$$

PO= the Proportion/Percentage of Observed agreement (as defined earlier)

PC= the Proportion/Percentage of Chance agreement (as also defined earlier)

SEM= the Standard Error of the Mean Difference between observed and expected agreement (PO-PC), calculated as the (Standard Deviation of the Mean Differences)/ $N^{1/2}$

where N refers to the Number of rater pairings

$N^{1/2}$= the square root of N and Z is interpreted in the standard manner, *via* two-tailed probability levels (*p*), as follows:

Value of Z:	*p* Value:
< 1.96	Not Statistically Significant (NS)
1.96	0.05
2.24	0.025
2.58	0.01
3.00	0.003
4.00	0.0001
5.00	<0.000001

*Signifies that for Ns as small as 20, t and Z are very similar, at 2.09 and 1.96 (essentially 2) at a two-tailed probability level of 0.05.

PO in the Context of Multiple Raters Examining a Single Case: Continuous-Ordinal (CO) Scales

The issue of defining the overall level of inter-rater agreement when multiple judges evaluate a single case involves many steps, but is not difficult to comprehend for non-biostatisticians. The steps are, as follows:

1. Noting whether the rating scale is based upon nominal or ordinal variables, or an admixture thereof.

2. Selecting an appropriate rater agreement paradigm that adequately classifies a given rating scale, as based upon the number of categories on the scale. It should be noted here that when the variables of study derive from a nominal scale, agreement is either 100% or 0%, *by definition*. However, when the variables of interest are based upon ordinal scales, there are levels of partial agreement that must also be considered. The number of such categories will, of course, vary, but it will always be equal to the number of rating categories minus 2 (the 100% agreement and 0% agreement categories).

PO in the Context of Multiple Raters Examining a Single Case: Dichotomous-Ordinal (DO) Scales

As introduced by Cicchetti (1976) [22], when the ordinal scale has a point of absence of a given diagnostic category, the number of agreement categories is equal to the number of scale points, k,

plus the number of intermediate or middle, m categories.

As an example in the field of medicine, consider a tumor classification system that is graded, by specific and non-overlapping clinical criteria as: Benign, Stage 1, Stage 2, Stage 3, or Stage 4 of the disease. It becomes intuitively obvious that a rater disagreement between Benign and Stage 1 is certainly clinically more serious than a rater disagreement between Stage 1 and Stage 2 of the disease. This means that a rational rater agreement system will award a lower rater agreement level to a Benign vs. Stage 1 disagreement than to one between Stage 1 and Stage 2, or between Stage 2 and Stage 3. Because of the "absence" category and two or more degree of "presence" categories, the scale was named dichotomous-ordinal (DO).

There are three possible types of rater agreement systems; linear, quadratic, and clinically relevant weights that are selected by the research investigator to best typify the relative seriousness of a given rater disagreement. Each of these rating systems will be discussed in turn.

Selecting the Type of Rater Agreement Weighting System: Linear Continuous-Ordinal (CO)

The Continuous-Ordinal (CO) rating scale is an ordinal scale as one typically conceives of one. As such, its focus is a graded system of increasing intensity of the "presence" of a particular phenomenon. Thus, one may think of a measurement of systolic blood pressure, with specific ranges defining, say 1="low", 2="average", 3= "high" blood pressure readings. What characterizes the CO scale is that rater disagreements the same number of categories apart can be considered clinically equivalent. This translates to mean that disagreements between "low" and "average" should typically be treated as clinically equivalent to rater discrepancies between "average" and "high."

The specific weights, for a given CO scale can be determined by the following simple formula, where k refers to the number of scale categories:

Linear Rater Partial Agreement Weights- CO Scales

These weights can be derived from the general formula:

$$Weights_{LINEAR} = k\text{-}1/k\text{-}1,\ k\text{-}2/k\text{-}1,\ k\text{-}3/k\text{-}1 \ldots k\text{-}k/k\text{-}1 \quad (4)$$

Table **1** - as adapted and summarized from Tables **1** and **2** of Cicchetti & Sparrow (1981), (pages 131 and 132) [17] shows linear agreement weights for CO ordinal scales ranging between 3 and 10 categories of classification.

Quadratic Rater Partial Agreement Weights-CO Scales

These weights are described in Fleiss & Cohen, 1973 [23]; Cicchetti & Sparrow, 1981 [17]: and more recently by Fleiss, Levin, & Cho-Paik, 2003 (p. 609) [16]. A given quadratic weight for any given CO scale, can be obtained by applying the following formula:

$$W_{ij} = 1 - (i - j)^2 / (k-1)^2 \quad (5)$$

where,

W refers to a given quadratic rater agreement weight,

i denotes the first rater in a given rater pairing,

j refers to the second rater in that pairing, and

Table 1: CO Linear Agreement Weights for 3 to 10 Category Ordinal Scales

	Number of categories separating a pair of ratings:							
k	**[ONE]**							
3	.50	**[TWO]**						
4	.67	.33	**[THREE]**					
5	.75	.50	.25	**[FOUR]**				
6	.80	.60	.40	.20	**[FIVE]**			
7	.83	.67	.50	.33	.17	**[SIX]**		
8	.86	.71	.57	.43	29	.14	**[SEVEN]**	
9	.875	.75	.625	.50	.375	.25	.125	**[EIGHT]**
10	.89	.78	.67	.56	.44	.33	.22	.11

Table 2: CO Quadratic Partial Agreement Weights for 3 to 10 Category Ordinal Scales

	Number of categories separating a pair of ratings:							
k	**[ONE]**							
3	.75	**[TWO]**						
4	.89	.56	**[THREE]**					
5	.94	.75	.44	**[FOUR]**				
6	.96	.84	.64	.36	**[FIVE]**			
7	.97	.89	.75	.56	.31	**[SIX]**		
8	.98	.92	.82	.67	.49	.27	**[SEVEN]**	
9	.98	.94	.86	.75	.61	.44	.23	**[EIGHT]**
10	.99	.95	.89	.80	.69	.56	.40	.21

k, as previously, refers to the number of categories on a particular DO scale

Quadratic agreement weights for CO scales ranging between 3 and 10 categories are spread in Table **2**.

Linear Rater Partial Agreement Weights- DO Scales

DO linear rater partial agreement weights are shown in Table **3**, for ordinal scales ranging between 3 and 10 categories of classification.

Quadratic Rater Partial Agreement Weights-DO Scales

DO quadratic rater partial agreement weights are given in Table **4**, for ordinal scales ranging between 3 and 10 categories of classification.

Comparing Linear and Quadratic Rater Weighting Systems

In comparing a linear and quadratic partial agreement weighting system, it becomes clear that: first, quadratic weights are consistently higher than their corresponding linear weights; second, linear weights distribute themselves more evenly than do quadratic weights; and third, linear weights seem to have, from another important perspective, a higher level of clinical intuitive appeal than do their quadratic counterparts. Specifically, Fleiss, Levin, & Cho-Paik (2003, p. 608) [16] support the use of linear rater agreement weights as "rational on clinical grounds." These points can be made more explicit by contrasting linear weights, Tables **1** and **3**, with their corresponding quadratic weights in Tables **2** and **4**.

Concerning the first point, it is always true that each partial agreement weight is higher when based upon a quadratic weighting system than when based upon a linear one. Concerning the second point, the quadratic weights distribute themselves unevenly, while the linear weights distribute very evenly. The latter are constructed so that the difference between each partial agreement weight is a constant. Thus, the partial linear weights for a six point CO scale are: 0.80, 0.60, 0.40,

Table 3: DO Linear Partial Agreement Weights for 3 to 10 k Category Ordinal Scales

	[ONE]															
k[1]	**CO[2]**	**DO[3]**	**[TWO]**													
3	.67	.33	**CO**	**DO**	**[THREE]**											
4	.80	.60	.40	.20	**CO**	**DO**	**[FOUR]**									
5	.86	.71	.57	.43	.29	.14	**CO**	**DO**	**[FIVE]**							
6	.89	.78	.67	.56	.44	.33	.22	.11	**CO**	**DO**	**[SIX]**					
7	.91	.82	.73	.64	.55	.45	.36	.27	.18	.09	**CO**	**DO**	**[SEVEN]**			
8	.92	.85	.77	.69	.62	.54	.46	.38	.31	.23	.15	.08	**CO**	**DO**	**[EIGHT]**	
9	.93	.87	.80	.73	.67	.60	.53	.47	.40	.33	.27	.20	.13	.07	CO	**DO**
10	.94	.88	.82	.76	.71	.65	.59	.53	.47	.41	.35	.29	.24	.18	.12	.06

1 indicates the number of ordinal categories; 2 designates a paired rater disagreement between degrees of "presence" of a given entity; and 3 refers to a paired rater disagreement between the " absence" and presence" of a given entity.

Table 4: Quadratic Partial Agreement Weights for 3 to 10 Category Ordinal Scales As a Function of the Number of Disagreement Categories Separating a Pair of Ratings; and the Construction of the Rating Scale- CO or DO

	[ONE]															
k[1]	CO[2]	DO[3]	[TWO]													
3	.89	.56	CO	DO	[THREE]											
4	.96	.84	.64	.36	CO	DO	[FOUR]									
5	.98	.92	.82	.67	.49	.27	CO	DO	[FIVE]							
6	.99	.95	.89	.80	.69	.56	.40	.21	CO	DO	[SIX]					
7	.99	.97	.93	.87	.79	.70	.60	.47	.33	.17	CO	DO	[SEVEN]			
8	.99	.98	.95	.91	.85	.79	.71	.62	.52	.41	.32	.15	CO	DO	[EIGHT]	
9	1.00	.98	.96	.93	.89	.84	.78	.72	.64	.56	.46	.36	.25	.13	CO	DO
10	1.00	.99	.97	.94	.91	.88	.83	.78	.72	.65	.58	.50	.42	.32	.22	.11

1 indicates the number of ordinal categories; 2 designates a paired rater disagreement between degrees of "presence" of a given entity; and 3 refers to a paired rater disagreement between the "presence" and "absence" of a given entity.

and 0.20, for inter-rater disagreements that are, respectively, 1, 2, 3, and 4 scale categories apart. Each increasing level of disagreement is separated by the same amount, or 0.20. Contrast this with the corresponding quadratic weights for the same six category CO scale, where the partial agreement weights are consistently higher at: 0.96, 0.84, 0.64, and 0.36. Note also the unevenness with which the weights distribute themselves. There is only a 0.04 discrepancy between Perfect agreement and a 1 category rater disagreement. Quite inconsistently, the difference between a 1 and 2 category rater discrepancy is 0.12; between a 2 and 3 category disagreement, the difference becomes 0.20, and between 3 and 4 categories of disagreement, the difference reaches a level of 0.28. These comparative examples lend weight to the argument that linear weights have more clinically-intuitive appeal than do quadratic weights, at least in the area of assessing levels of inter-examiner agreement. These weaknesses of quadratic rater partial agreement weights also become progressively more problematic as the number of ordinal scale categories increases. As shown in Table **4**, this reaches its zenith when the number of ordinal categories reaches 9 or 10. For both these scales, a one category disagreement results in a nonsensical perfect partial agreement weight of 1.00; and for the latter, two category and three category disagreements receive respective weights of 0.99, 0.97, 0.94, and 0.91. Contrast these with their linear weight counterparts in Table **3**, of 0.94, 0.88, 0.82, and 0.76, respectively. The latter are congruent with the aforementioned Fleiss, *et al*. (2003, p.608) [16] designation of linear weighting systems as being "rational on clinical grounds."

All this said, it is also recognized that quadratic functions have been integrally related to many major discoveries in mathematics, such as the familiar Pythagorian theorem, whereby, in a right-angled triangle, the square of the hypotenuse (h) is equal to the sums of the squares of the two sides of the triangle, a and c, such that $h^2 = a^2 + c^2$. In a much broader sense, the Analysis of Variance (ANOVA) itself is based upon quadratic principles, with its focus on sums of squares. The point here is that while quadratic weighting systems are not ideal for the assessment of inter-rater agreement, they are nonetheless most appropriate in other areas of statistical and mathematical assessment.

Clinician Derived Rater Agreement Weights

For some diagnostic issues, neither linear nor quadratic systems are adequate to define clinically meaningful rater partial agreement weights**.** Here the clinical researcher is motivated to devise her/his own system of weights.

As one example, the MRI diagnosis of the size of the human hippocampus can be classified into five ordinal categories, based upon a set of well-defined, non-overlapping categories of classification, such that: 1=definitely normal; 2=probably normal; 3=equivocal; 4=probably abnormal; and 5=definitely abnormal. The Yale epileptologist Professor Rick Bronen defined the partial agreement weights for this clinical scale as the following: 1-2, 2-1, 4-5, and 5-4 disagreements received a weight of 0.90; 2-3, 3-2, 3-4, and 4-3 discrepancies were given a weight of 0.80; 1-3, 3-1, 3-5, and 5-3 pairings received a weight of 0.50; 2-4 and 4-2 disagreements were given a weight of 0.20; and 1-

4, 4-1, 2-5, and 5-2 pairings received a weight of 0.10 [24].

In the upcoming sections of the report, the authors will demonstrate how the statistic can be utilized; and the new methodology will next be illustrated for the assessment of PTSD, when multiple examiners evaluate a Vietnam era patient.

Multiple Examiners Evaluate a Single Case: Establishing Levels of Agreement

As an example, one of the items from the Clinician Administered PTSD Scale (CAPS-1), due to Blake, *et al*. (1995) [1] defines both the *frequency* and the *intensity* of a wide range of PTSD symptoms. One of the items refers to the *frequency* of "difficulty falling asleep". The clinician queries a given patient by prompting, as follows:

How often in the past month have you had difficulty falling asleep?

0=Never

1=Once or twice

2=Once or twice a week

3=Several times a week

4=Daily or almost every day

In scoring the *intensity* of the same symptom, the clinician prompts are defined, as follows:

How much effort did you make to avoid difficulty in falling asleep?

0=No effort

1=Mild, minimal effort

Table 5: Item-by-Item Reliability of the CAPS-1 *Frequency of* Symptomatology: First Patient[1]

	Significance Levels:			
	PO	**K_w**	**Clinical**	**Statistical**
A. By Criterion B, whereby the patient's traumatic event was persistently re-experienced as				
1. Recurrent and intrusive recollections	.98	.93	Excellent	< 0.0001
2. Distress when exposed to the event	1.00	1.00	Perfect	< 0.0001
3. Acting or feeling as if the event was recurring	1.00	1.00	Perfect	<0.0001
4. Recurring distressing dreams of the event	.93	.77	Excellent	<0.0001
B. By Criterion C, whereby the patient demonstrated avoidance of stimuli, numbing of responsiveness, as characterized by				
5. Efforts to avoid thoughts or feelings	1.00	1.00	Perfect	<0.0001
6. Efforts to avoid activities or situations	1.00	1.00	Perfect	<0.0001
7. Inability to recall trauma aspects	1.00	1.00	Perfect	<0.0001
8. Markedly diminished interest in activities	1.00	1.00	Perfect	<0.0001
9. Feelings of detachment or estrangement	1.00	1.00	Perfect	<0.0001
10. Restricted range of affect	1.00	1.00	Perfect	<0.0001
11. A sense of a foreshortened future	1.00	1.00	Perfect	<0.0001
C. By Criterion D, whereby the patient reported persistent symptoms of increased arousal, namely				
12. Difficulty in falling or staying asleep	.93	.77	Excellent	<0.0001
13. Irritability or outbursts of anger	1.00	1.00	Perfect	<0.0001
14. Difficulty in concentrating	1.00	1.00	Perfect	<0.0001
15. Hypervigilance	1.00	1.00	Perfect	<0.0001
16. An exaggerated startle response	1.00	1.00	Perfect	<0.0001
17. Physiologic reactivity	.88	.60	Good	<0.0001
D. By Criterion E, whereby the patient reported associated features of guilt, in terms of.				
18. Guilt over certain acts	1.00	1.00	Perfect	<0.0001
19. A reporting of survival guilt	.94	.80	Excellent	<0.0001

[1]The Four Composite Scores, Averaged Over the Items Comprising Criteria B Through E, As Well As the Global Score, Averaged Across the 19 Symptoms, Reached 100% Agreement Across the 12 Examiners.

2=Moderate, some effort, avoidance definitely present

3-Severe, considerable effort, marked avoidance

4= Extreme, drastic attempts at avoidance

Each of the PTSD symptoms is presented in Table **5**.

The steps for assessing inter-examiner agreement when multiple examiners evaluate a single case can be summarized as the following:

1. Based upon the type of clinical scale, select an appropriate rater partial agreement system.
2. Arrange the data preparatory to analysis.
3. Calculate weighted Kappa.
4. Assess level of clinical or practical significance.
5. Assess level of statistical significance.

Selecting an Appropriate Weighting System

As indicated, both the *frequency* and *intensity* of each PTSD symptom were assessed on five category Dichotomous-Ordinal (DO) scales. The appropriate linear weighting system appears in Table **3**, producing the following rater partial agreement weights: .86, .71, .57, .43, .29, and .14.

RESULTS AND DISCUSSION

Determining Levels of Clinical Significance

The number of paired comparisons, for a given CAPS-1 item, is calculated using the formula E (E-1)/2,

Table 6: Item-by-Item Reliability of the CAPS-1 *Intensity of* Symptomatology: First Patient[1]

	Significance Levels:			
	PO	**K_w**	**Clinical**	**Statistical**
A. By Criterion B, whereby the patient's traumatic event was persistently re-experienced as				
1. Recurrent and intrusive recollections	.93	.77	Excellent	< 0.0001
2. Distress when exposed to the event	.96	.87	Excellent	< 0.0001
3. Acting or feeling as if the event was recurring	.98	.93	Excellent	<0.0001
4. Recurring distressing dreams of the event	1.00	1.00	Excellent	<0.0001
B. By Criterion C, whereby the patient demonstrated avoidance of stimuli, numbing of responsiveness, as characterized by				
5. Efforts to avoid thoughts or feelings	.96	.87	Perfect	<0.0001
6. Efforts to avoid activities or situations	1.00	1.00	Perfect	<0.0001
7. Inability to recall trauma aspects	1.00	1.00	Perfect	<0.0001
8. Markedly diminished interest in activities	.92	.73	Good	<0.0001
9. Feelings of detachment or estrangement	.98	.93	Excellent	<0.0001
10. Restricted range of affect	.98	.93	Excellent	<0.0001
11. A sense of a foreshortened future	.95	.83	Excellent	<0.0001
C. By Criterion D, whereby the patient reported persistent symptoms of increased arousal, namely				
12. Difficulty in falling or staying asleep	.90	.67	Good	<0.0001
13. Irritability or outbursts of anger	1.00	1.00	Perfect	<0.0001
14. Difficulty in concentrating	.93	.77	Excellent	<0.0001
15. Hypervigilance	.96	.87	Excellent	<0.0001
16. An exaggerated startle response	1.00	1.00	Perfect	<0.0001
17. Physiologic reactivity	.98	.93	Excellent	<0.0001
D. By Criterion E, whereby the patient reported associated features of guilt, in terms of.				
18. Guilt over certain acts	.94	.80	Excellent	<0.0001
19. A reporting of survival guilt	1.00	1.00	Perfect	<0.0001

[1]The Four Composite Scores, Averaged Over the Items Comprising Criteria B Through E, As Well As the Global Score, Averaged Across the 19 Symptoms, Reached 100% Agreement Across the 12 Examiners.

where E refers to the number of Examiners. For these data, there are (12 x 11)/2 or 66 pairings.

With regard to the first Veteran patient (Table **5**) there was 100% agreement among the 12 Examiners on 14 of the 19 *frequency* of PTSD symptomatology items. The remaining items showed PO values ranging between 88%, with a K_w value of 0.60 (Good agreement), and 98%, with a corresponding K_w of 0.93 (Excellent agreement). All 5 Composite scores, based upon averaging the items within Criteria B, C D, and E, and the Global score, each manifested 100% agreement across the 12 Examiners.

Concerning the *intensity* of PTSD symptomatology, the data in Table **6** indicate that 6 of the items revealed a perfect reliability level of 100%. The remaining 13 items evidenced PO values ranging between 90% and 98%, with corresponding K_w values of .67 (Good) and .93 (Excellent).

With respect to *frequency* of PTSD symptomatology, 13 of the 19 items evidenced 100% agreement. Of the remaining 6 items, item 12 manifested a level of Good agreement, with a PO of 88% and a corresponding K_w of 0.60. The reliability levels of the remaining 5 items ranged between 92% and 98%, with respective K_w levels of 0.73 (Good) and 0.93 (Excellent).

The results for the second cross-validation patient are given in Tables **7** and **8**, first, for the *frequency* of PTSD symptomatology, then for the *intensity* of PTSD symptomatology.

The results for the *frequency* of PTSD symptomatology, for the second patient are shown in

Table 7: Item-by-Item Reliability of the CAPS-1 *Frequency of* Symptomatology: Second Patient[1]

	Significance Levels:			
	PO	**K_w**	**Clinical**	**Statistical**
A. By Criterion B, whereby the patient's traumatic event was persistently re-experienced as				
1. Recurrent and intrusive recollections	1.00	1.00	Perfect	< 0.0001
2. Distress when exposed to the event	1.00	1.00	Perfect	< 0.0001
3. Acting or feeling as if the event was recurring	1.00	1.00	Perfect	<0.0001
4. Recurring distressing dreams of the event	1.00	1. 00	Perfect	<0.0001
B. By Criterion C, whereby the patient demonstrated avoidance of stimuli, numbing of responsiveness as characterized by				
5. Efforts to avoid thoughts or feelings	1.00	1.00	Perfect	<0.0001
6. Efforts to avoid activities or situations	.96	.86	Perfect	<0.0001
7. Inability to recall trauma aspects	.95	.84	Excellent	<0.0001
8. Markedly diminished interest in activities	1.00	1.00	Perfect	<0.0001
9. Feelings of detachment or estrangement	1.00	1.00	Perfect	<0.0001
10. Restricted range of affect	1.00	1.00	Perfect	<0.0001
11. A sense of a foreshortened future	1.00	1.00	Perfect	<0.0001
C. By Criterion D, whereby the patient reported persistent symptoms of increased arousal, namely				
12. Difficulty in falling or staying asleep	.88	.60	Good	<0.0001
13. Irritability or outbursts of anger	1.00	1.00	Perfect	<0.0001
14. Difficulty in concentrating	.93	.77	Excellent	<0.0001
15. Hypervigilance	.92	.73	Good	<0.0001
16. An exaggerated startle response	1.00	1.00	Perfect	<0.0001
17. Physiologic reactivity	.98	.93	Excellent	<0.0001
D. By Criterion E, whereby the patient reported associated features of guilt, in terms of.				
18. Guilt over certain acts	1.00	1.00	Perfect	<0.0001
19. A reporting of survival guilt	1.00	1.00	Perfect	<0.0001

[1]The Four Composite Scores, Averaged Over the Items Comprising Criteria B through E, as Well As the Global Score, Averaged Across the 19 Symptoms, Ranged Between 92% and 100% Agreement Across the 12 Examiners.

Table 8: Item-by-Item Reliability of the CAPS-1 *Intensity of* Symptomatology: Second Patient[1]

	Significance Levels:			
	PO	**K_w**	**Clinical**	**Statistical**
A. By Criterion B, whereby the patient's traumatic event was persistently re-experienced as				
1. Recurrent and intrusive recollections	.96	.87	Excellent	< 0.0001
2. Distress when exposed to the event	1.00	1.00	Perfect	< 0.0001
3. Acting or feeling as if the event was recurring	1.00	1.00	Perfect	<0.0001
4. Recurring distressing dreams of the event	1.00	1. 00	Perfect	<0.0001
B. By Criterion C, whereby the patient demonstrated avoidance of stimuli, numbing of responsiveness as characterized by				
5. Efforts to avoid thoughts or feelings	.94	.80	Excellent	<0.0001
6. Efforts to avoid activities or situations	.94	.80	Excellent	<0.0001
7. Inability to recall trauma aspects	.95	.83	Excellent	<0.0001
8. Markedly diminished interest in activities	.93	.77	Excellent	<0.0001
9. Feelings of detachment or estrangement	1.00	1.00	Perfect	<0.0001
10. Restricted range of affect	.98	.93	Excellent	<0.0001
11. A sense of a foreshortened future	.96	.87	Excellent	<0.0001
C. By Criterion D, whereby the patient reported persistent symptoms of increased arousal, namely				
12. Difficulty in falling or staying asleep	.85	.50	Good	<0.0001
13. Irritability or outbursts of anger	1.00	1.00	Perfect	<0.0001
14. Difficulty in concentrating	.93	.77	Excellent	<0.0001
15. Hypervigilance	.92	.73	Good	<0.0001
16. An exaggerated startle response	.94	.80	Excellent	<0.0001
17. Physiologic reactivity	.96	.87	Excellent	<0.0001
D. By Criterion E, whereby the patient reported associated features of guilt, in terms of.				
18. Guilt over certain acts	.97	.90	Excellent	<0.0001
19. A reporting of survival guilt	.97	.90	Excellent	<0.0001

[1]The Four Composite Scores, Averaged Over the Items Comprising Criteria B through E, as Well As the Global Score, Averaged Across the 19 Symptoms, Reached 100% Agreement Across the 12 Examiners.

Table 7. There was Perfect agreement on 13 of the PTSD symptoms. The remaining 6 symptoms demonstrated reliability levels ranging between 88%, with a chance-corrected level of 0.60 (Good agreement) and 98% (Excellent agreement), with a chance-corrected level of 0.93 (Excellent agreement). The Composite Scores ranged between 92% and 100%.

Finally, the findings for the *intensity* of PTSD symptomatology in the cross-validation patient, indicated that there was 100% agreement, among the 12 Examiners, on 5 of the 19 PTSD symptoms. For the remaining symptoms reliability levels ranged between 85% and 98%, with respective K_w levels of .50 (Fair) and .93 (Excellent). All 5 Composite scores demonstrated 100 % agreement.

Determining Levels of Statistical Significance

We shall utilize the results that occurred when the 12 clinicians evaluated the responses of the second patient to Item 8 (Table **8**): *Intensity* of *Markedly diminished interest in activities*. As shown below, the (12 X 11)/2, or 66 rater pairings distributed themselves, as follows: 11 raters gave a score of 0 and the remaining one a score of 1. The 66 rater pairings distributed themselves as: PO= [(55 X 1) + (11 X .57)/66= 93%. This produced a K_w value of (.93-.70)/.30= 0.77, an Excellent chance-corrected overall level of agreement by the criteria of Cicchetti & Sparrow (1981) [17], and Substantial, by the conceptually similar criteria of Landis & Koch (1977) [14]. To calculate the level of statistical significance, we apply formula 3 to obtain:

Z= (PO-PC)/SEM= (.93 -.70)/.010= 23, thereby producing $p< 0.00001$.

While this overall level of reliability across the 12 clinicians is quite impressive, it does not inform as to whether any of the 12 judges varied appreciably from the remaining ones. This information can be easily

obtained by referring once more to the data obtained for the overall agreement level of 93%, a few paragraphs ago. Note that 11 of the raters (55 pairings) were in 100% agreement (all with a 0 rating); but the remaining rater was awarded a score of 1, producing an agreement level of 0.57 with each of the remaining judges. This produces a K_w value of only (.57-.70)/.30= - 0.43, which is far below statistical or clinical significance. The importance of this finding is that a very high level of clinical and statistical inter-rater agreement (here 93%) can mask the fact that at least one of the raters deviates greatly from the overall level of agreement.

In a broader sense, these two examples serve to support the aforementioned distinction and relationship between nomothetic and idiographic approaches to scientific research, with the former exemplified by overall inter-judge agreement and the latter by the specific levels of agreement for each individual judge. They also highlight the necessity for applying both approaches whenever possible. Finally, they raise the question of whether the low agreement levels of one or more raters on any given PTSD item (1-24), patient (1 or 2), component of symptomatology (*Frequency or Intensity*), or number of raters (12), serves to indicate scoring bias.

Testing for Rater Bias: *Overall* Agreement Is Acceptable, But Individual Rater Agreement is Not

The total number of overall rater scorings amounts to: (24 items X 2 types of symptomatology X 2 patients)= 96 scorings.

There were 8 instances in which the overall rater agreement (PO) was acceptable, but individual rater agreement was not: [(85% (Fair); 88% (Good); 90% (Good); 93% (Excellent)-2 items; and 95% (Excellent)-3 items., 4 of the 8 instances pertained to each of the two patients, indicating that there was no bias by patient. There was also no discernible bias by component of symptom, with 5 instances of ratings of the *Intensity* of PTSD symptomatology and 3 instances of ratings of the *Frequency* of PTSD symptomatology. There was also no pattern of bias by the actual item itself: Of the 8 instances of potential bias, by specific PTSD item, 3 pertained to item 7 (Inability to recall traumatic event); 3 pertained to item 12 (Difficulty falling or staying asleep); 1 pertained to item 4 (Recurring distressing dreams of the event); and the last one pertained to item 17 (Physiologic reactivity). Finally, there was no bias by raters. The 9 instances of unacceptable individual scorings distributed themselves widely, across the raters and PTSD symptoms, as follows:

Rater 2 (Item 7-Intensity; score of 71%- PO =95%);

Rater 3 (Item 4; score of 57%- PO=93%);

Rater 4 (Item 7; score of 71%- PO=95%);

Also Rater 4 (Item 12; score of 71%-PO=85%);

Rater 5 (Item 12; score of 57%- PO=93%);

Rater 6 (Item17; score of 29%- PO=88%);

Rater 11 (Item 7; score of 71%; PO=95%);

Rater 11 (Item 12; score of 71%; PO=85%); and

Rater 12 (Item score of 62%-PO=90%).

BRIEF SUMMARY AND CONCLUSIONS

The purpose of this research was to assess the statistical and clinical significance of the reliability ratings of both the *frequency* and *intensity* of PTSD symptomatology when multiple examiners evaluated a single case. In this situation, 12 examiners evaluated two Vietnam era patients, with the second patient serving cross-validation purposes. The data were also examined, for the first time, in terms of overall and specific rater reliability levels. Finally a procedure was developed for examining whether the results were affected by component of: PTSD symptomatology (*frequency-intensity*); item (1-24); or specific rater (1-12).

Results indicated that all items reached levels of statistical and clinical significance by accepted scientific standards. While there were no biases evident in these results, they did demonstrate that even when an overall reliability level is as high as 95%, an excellent level of chance-corrected agreement that is also highly statistically significant, it can and will, as was shown, mask the fact that one or more raters may be performing at levels that fail to meet accepted criteria for clinical and statistical significance. The value of this finding is that it has both medical and bio-behavioral Implications well beyond the study of the reliability of PTSD symptomatology, such as Autism [25]. It would also seem ideal for many areas of medical diagnosis such as the diagnostic decision as to whether a given lesion is Benign, Stage 1, Stage 2, Stage 3 or Stage 4.

REFERENCES

[1] Blake DD, Weathers FW, Nagy LM, Kaloupek DG, Gusman FD, Charney DS, *et al.* The development of a clinician administered PTSD scale (CAPS). J Traumatic Stress 1995; 8: 75-90. http://dx.doi.org/10.1002/jts.2490080106

[2] Cicchetti DV, Fontana A, Showalter, D. Evaluating the reliability of multiple assessments of PTSD symptomatology: Multiple examiners, one patient. Psychiat Res 2009; 166: 269-280. http://dx.doi.org/10.1016/j.psychres.2008.01.014

[3] Windelband W, Oakes G. History and natural science. Hist Theory 1894/1980; 19: 165-168.

[4] Windelband W. A history of philosophy 1901/2001; New Jersey: Paper Tiger.

[5] Allport G. The functional autonomy of motives. Am J Psychol 1937; 50: 141-156. http://dx.doi.org/10.2307/1416626

[6] Cicchetti DV. On the psychometrics of neuropsychological measurement: A biostatistical perspective. Oxford handbook of neuropsychology, New York, NY: Oxford University Press 1989.

[7] Robinson OC. The ideographic/nomothetic dichotomy: Tracing historical origins of contemporary confusions. Hist Phil Psychol 2011; 13: 32-39.

[8] Holschu N. Randomization and design I. In: Fisher RA: An appreciation. Fienberg SE, Hinkley DV, Eds. Lecture notes in statistics. New York, NY: Springer 1980; pp. 35-45.

[9] Kelly K, Preacher KJ. On effect size. Psychol Meth 2012; 17: 137-152. http://dx.doi.org/10.1037/a0028086

[10] Cohen J. Statistical power analysis for the behavioral sciences. Glendale, NJ: Lawrence Erlbaum Associates 1988.

[11] Borenstein M. The shift from significance testing to effect size estimation. In: Bellak AS, Hershen M, series Eds., and Schooler N, volume editor, Research and methods: Comprehensive clinical psychology. New York, NY: Pergammon 1998; volume 3: pp. 313-349.

[12] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 23: 37-46. http://dx.doi.org/10.1177/001316446002000104

[13] Cohen J. Weighted kappa: Nominal scale agreement with provision for partial credit. Psychol Bull 1968; 70: 213-220. http://dx.doi.org/10.1037/h0026256

[14] Landis JR, Koch GG. The measurement of agreement for categorical data. Biometrics 1977; 33: 159-174. http://dx.doi.org/10.2307/2529310

[15] Fleiss J. Statistical methods for rates and proportions. New York, NY: Wiley (2nd ed.) 1981.

[16] Fleiss J, Levin B, Cho-Paik M. Statistical methods for rates and proportions. New York, NY: Wiley (3rd ed.) 2003. http://dx.doi.org/10.1002/0471445428

[17] Cicchetti DV, Sparrow SS. Developing criteria for establishing inter-rater reliability of specific items: Applications to assessment of adaptive behavior. Am J Mental Deficiency 1981; 86: 127-137.

[18] Cicchetti DV, Volkmar F, Klin A, Showalter D. Diagnosing autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. Child Neuropsychol 1995; 1: 26-37.

[19] Cicchetti DV, Bronen R, Spencer S, Haut S, Berg A, Oliver P, Tyrer P. Rating scales, scales of measurement, issues of reliability: Resolving some critical issues for clinicians and researchers. J Nervous Mental Disease 2006; 194: 557-564.

[20] Szalai, JP. The statistics of agreement on a single item or object by multiple raters. Percept Motor Skills 1993; 77: 377-378. http://dx.doi.org/10.2466/pms.1993.77.2.377

[21] Szalai, JP. Kappa SC. A measure of agreement on a single rating category for a single item or object rated by multiple raters. Psychol Reports 1998; 82: 1321-1322. http://dx.doi.org/10.2466/pr0.1998.82.3c.1321

[22] Cicchetti DV. Assessing inter-rater reliability for rating scales: Resolving some basic issues. Br J Psychiat 1976; 12: 452-456. http://dx.doi.org/10.1192/bjp.129.5.452

[23] Fleiss J, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Pschol Meas 1973; 33: 613-619. http://dx.doi.org/10.1177/001316447303300309

[24] Bronen RA, Chan S, Cicchetti DV, Berg AT, Spencer S. Inter-rater agreement for MRI interpretation of epilepsy surgery patients. Am Epilep Soc 2004.

[25] Cicchetti, DV, Lord C, Koenig K, Klin A, Volkmar FR. Reliability of the ADI-R for the single case-Part II: Clinical versus statistical significance. J Autism Develop Disord 2014; 44: 3154-3160.

Received on 06-09-2014 Accepted on 27-11-2014 Published on 27-01-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.01.4