Optimal Data Analysis
Vol. 3 (April 13, 2014), 42-49

2155-0182/10/$3.00

How to Assess Inter-Observer Reliability of Ratings Made on Ordinal Scales: Evaluating and Comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

An exact, optimal (“maximum-accuracy”) psychometric methodology for assessing inter-observer reliability for measures involving ordinal ratings is used to evaluate and compare two emergency medicine triage algorithms—both of which classify patients into one of five ordinal categories. Ten raters independently evaluated the identical set of 200 patients, five with each algorithm. Analysis revealed moderate levels of inter-observer reliability, indicating that prior estimates of almost perfect inter-observer reliability obtained for the present data using suboptimal statistical methods are untenable.

Ordinal ratings are ubiquitous in modern society, including—appropriately so—within the scientific literature. Examples of well-known ordinal ratings are Homeland Security alert status, report-card grade, and disease stage. In general, data for all phenomena measured on categorical ordinal, or on Likert-type scales constitute ordinal ratings.[1-3]

In spite of their omnipresence, ordinal data are difficult to analyze appropriately vis-à-vis conventional statistical means. Parametric methods such as Student’s *t*-test and analysis of variance, or Pearson correlation and multiple regression analysis, were developed for use with interval data, and sample data are assumed to be distributed in a specific manner as well as to fulfill other requirements.[4-7] It thus isn’t entirely surprising that a relatively common error is to treat ordinal data as though they are assessed on a categorical scale, and employ chi-square or logistic regression in statistical analysis of the data.[8-10] A conceptually related error involves arbitrarily parsing an ordinal scale into a set of “categorical” partitions (eyeball strata) in order to facilitate nominal statistical analysis: such arbitrary parsing can mask effects present in the unaltered raw data.[11-13] Finally, another nominal statistic, kappa, is used in the analysis of ordinal data, sometimes involving arbitrarily and non-normed weighting schemes.[14-17]

The conundrum of what method to use in statistical analysis of ordinal data has recently come into focus in emergency medicine. The issue is how to assess inter-observer (or inter-rater) reliability of ratings produced by two or more independent nurses, both of whom are using an algorithm to assign triage scores (which may assume one of five possible ordinal values) to patients arriving at the emergency department (ED). Kappa[18] and quadratically-weighted kappa[19] estimated the inter-observer reliability of these triage scores to be nearly perfect (kappa and intraclass correlation[20] were 0.89 or greater). Of course, these findings have been criticized on the basis of well-known problems underlying these methodologies.[21,22]

In contrast to intractably problematic suboptimal methods, UniODA is ideally-suited to conduct exact, optimal (maximum-accuracy) statistical analysis for data involving measurements made using ordinal scales.[1,2,8-10] The use of UniODA to assess inter-observer, test-retest and parallel-forms reliability, and in structural decomposition (which is analogous to principal components analysis except that accuracy is explicitly maximized, instead of variance[23]) of ordinal reliability data, is well-documented.[1,15,16] In the present application UniODA is used to test the *a priori* (directional) hypothesis that triage codes (integers ranging from 1 to 5), which were assigned to a sample of patients by two independent nurses, are consistent. This *a priori* hypothesis is defined in UniODA (by the *directional* command[1]) as shown in Figure 1.

Figure 1: Illustration of UniODA *a priori* Hypothesis that Triage Codes for a Pair of Nurse Raters Agree

Rater-A		Rater-B
1	→	1
2	→	2
3	→	3
4	→	4
5	→	5

As illustrated, if Rater-A assigns a triage code of 1 to an ED patient, the UniODA model predicts Rater-B likewise assigns a triage code of 1 to the patient; if Rater-A assigns a triage code of 2 to a patient, the model predicts Rater-B likewise assigns a triage code of 2 to the patient; and so forth.

Methods

Data in the present study were analyzed previously using quadratically-weighted kappa[19] and are reanalyzed here[24] using UniODA[1] run on MegaODA software.[25-27]

Described[19] in the original analysis, ten triage nurses, all previously trained in use of the 5-point *C*anadian Emergency Department *T*riage and *A*cuity *S*cale (CTAS) were randomized into one of two conditions. Five nurses (Rater-1 through Rater-5) were trained in use of the 5-point *E*mergency *S*everity *I*ndex (ESI) Version 3 triage algorithm. The other five nurses (Rater-6 through Rater-10) were instead given refresher training in the CTAS triage algorithm. All training sessions required three hours. Each nurse independently assigned triage scores using either the ESI or the CTAS to each of "…200 case scenarios abstracted from prospectively collected, local ED cases" (p. 242).

Evaluating the Inter-Observer Reliability of Scores on the ESI

First, separately for each unique rater pairing, UniODA was used to evaluate the *a priori* hypothesis that the triage codes of the raters agree (Figure 1). The *a priori* UniODA model was consistent with the data of four of the total of ten unique rater pairings.

Strongest inter-observer reliability was obtained for the (Rater-1, Rater-2) pairing, for which overall agreement was 61.5%, and the *a priori* UniODA model achieved ESS=59.9 (p<0.0001), indicating relatively strong inter-observer agreement.[28] Consistent performance was obtained in jackknife validity analysis, so

Copyright 2014 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

these results are expected to cross-generalize to an independent random sample of ED patients.[1]

Second-strongest inter-observer reliability was obtained for the (Rater-1, Rater-3) pairing: overall agreement was 61.5%, and the *a priori* model achieved ESS=47.9 (*p*<0.0001), indicating moderate inter-observer agreement. In jackknife analysis overall agreement dropped to 61.0%, and ESS fell to 35.4: thus, diminished inter-observer reliability is expected if this pair assesses an independent random patient sample.

Third-strongest inter-observer reliability occurred for the (Rater-2, Rater-3) pairing for which overall agreement=59.5%, and the *a priori* model achieved ESS=45.3 (*p*<0.0001), reflecting moderate inter-observer agreement. Overall agreement dropped to 59.0%, and ESS fell to 37.0 in jackknife analysis.

Weakest inter-observer reliability was obtained for the (Rater-2, Rater-4) pairing, for which overall agreement was 57.8%, and the *a priori* model achieved ESS=39.4 (*p*<0.0001), indicating moderate inter-observer agreement. Overall agreement fell to 57.3%, and ESS fell to 31.1 in jackknife analysis.

For all six remaining parings the *a priori* model was untenable—no UniODA model was possible for the *a priori* hypothesis given the actual data. Thus, for these remaining pairings the *a priori* hypothesis was dropped, and an exploratory UniODA analysis was conducted.[1]

For two pairings the same exploratory UniODA model was identified, which involved all five possible triage codes (Figure 2). For the (Rater-1, Rater-4) pairing, overall agreement= 38.7%, ESS=36.6 (moderate inter-observer reliability; *p*<0.0001; stable in jackknife analysis). For the (Rater-3, Rater-4) pairing, overall agreement=38.2%, ESS=33.9 (moderate inter-observer reliability; *p*<0.0005; too few patients were rated as 1's to conduct jackknife analysis). As illustrated, if Rater-4 assigns a triage code of 1 to an ED patient, the UniODA model predicts Rater-1 and Rater-3 assign a triage code of 2 to the patient; if Rater-4 assigns a triage code of 2 to a patient, then the model predicts Rater-1 and Rater-3 assign a triage code of 1 to the patient; and so forth.

Figure 2: Exploratory UniODA Model for the (Rater-1, Rater-4) and (Rater-3, Rater-4) Pairings

Rater-4		Rater-1, Rater-3
1	→	2
2	→	1
3	→	3
4	→	4
5	→	5

This is a classic example of one of four general types of *reliable nonlinear patterns* that may underlie reliability data, which have been described[1] (specifically, Type A; pp. 137-138). ESI ratings by these two pairs demonstrate *local regression* at lower levels of the scale (across the two most severe triage codes), but are stable over the remaining range of the scale.

For the remaining four rater pairings—all of which involve Rater-5, no UniODA model was possible that included all five triage levels. Thus, exploratory UniODA was allowed to forego the use of one or more class categories (triage codes) in the model. This is known as a *degenerate solution*, and it is used to identify an optimal model in applications involving sparse or missing class categories.[1]

The strongest exploratory degenerate model was obtained for the (Rater-1, Rater-5) pairing: overall agreement=35.5%, ESS=32.8 (moderate agreement), *p*<0.0002, stable in jackknife analysis. Figure 3 illustrates the UniODA model. The model reveals collapsing (*local compression*) for the most serious cases (triage codes<3) for this pair, but consistent assignments for triage codes ≥3. The UniODA model is degenerate because no predictions of triage code 2 are made for Rater-1, Rater-3 or Rater-4 (see Figure 3).

Figure 3: Exploratory Degenerate UniODA Model for (Rater-1, Rater-5), (Rater-3, Rater-5), and (Rater-4, Rater-5) Pairings

Rater-5		Rater-1, Rater-3, Rater-4
1,2	→	1
3	→	3
4	→	4
5	→	5

The second-strongest exploratory degenerate model was obtained for the (Rater-2, Rater-5) pairing: overall agreement=25.5%, ESS=32.5 (moderate agreement), *p*<0.0001. Overall agreement fell to 22.0%, and ESS fell to 20.8 (relatively weak agreement) in jackknife analysis. The UniODA model is illustrated in Figure 4. There is local compression for this pair for the more serious cases, as well as for the less serious cases—indicating polarization. This is a Type D nonlinear reliability model[1] with the middle code uncompressed (p. 137).

Figure 4: Exploratory Degenerate UniODA Model for (Rater-2, Rater-5) Pairing

Rater-5		Rater-2
1,2	→	1
3	→	3
4,5	→	5

The third-strongest exploratory degenerate model occurred for the (Rater-3, Rater-5) pairing: overall agreement=34.0%, ESS=29.3 (moderate agreement), *p*<0.03 (while this is statistically significant at the generalized criterion, it *isn't* statistically significant at the experimentwise criterion[1]). The UniODA model is illustrated in Figure 3. Jackknife analysis was not possible as the required minimum of two observations per class category wasn't met.[1]

And finally, the weakest exploratory degenerate model was obtained for the (Rater-4, Rater-5) pairing: overall agreement=29.6%, ESS=28.6 (moderate agreement), *p*<0.05 (while this is statistically significant at the generalized criterion, it *isn't* statistically significant at the experimentwise criterion[1]). The UniODA model is illustrated in Figure 3. Jackknife analysis was not possible because of sparse data.

Table 1 summarizes the ESS achieved by the UniODA model for the training analyses involving the ESI triage ratings.

Table 1: ESI inter-observer ESS results

	Rater-2	Rater-3	Rater-4	Rater-5
Rater-1	59.9	47.9	38.7*	32.8**
Rater-2		45.3	39.4	32.5**
Rater-3			33.9*	29.3**
Rater-4				28.6**

Note: Tabled is ESS for training analysis predicting ratings of one rater given ratings of the other rater: ESS=0 is the level of agreement expected by chance; ESS=100 is perfect agreement. Entries marked by an asterisk were obtained by exploratory model; entries marked by two asterisks were obtained by exploratory degenerate model.

These findings clearly demonstrate that the inter-observer agreement observed for ESI scores is far from perfect. For only one of the ten rater pairs was agreement relatively strong, yielding 59.9% of the gain in agreement that it is theoretically possible to attain above what is expected by chance.[1] For only four of ten rater pairings was the *a priori* UniODA model even feasible, and the other six models indicated patterns of *consistent disagreement*. All but one of ten models achieved mediocre ESS, and six models identified reliable inconsistencies such as compression, omission, and regression. Results for two models weren't statistically significant at the experimentwise criterion.[1]
It is thus concluded that prior kappa-based and quadratically-weighted-kappa-based estimates, which suggested nearly perfect inter-observer reliability of ESI triage scores, are untenable.

Copyright 2014 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

Evaluating the Inter-Observer Reliability of Scores on the CTAS

As done for ESI triage data, separately for each unique rater pairing, UniODA was used to evaluate the *a priori* hypothesis that CTAS triage codes of the raters agree (Figure 1). The *a priori* UniODA model was consistent with the data of all ten unique rater pairings.

Strongest inter-observer reliability was obtained for the (Rater-7, Rater-9) pairing, for which overall agreement was 52.5%, and the *a priori* UniODA model achieved ESS=53.6 (*p*<0.0001), indicating relatively strong inter-observer agreement.[28] Consistent performance was obtained in jackknife analysis.

The second-strongest inter-observer reliability was obtained for the (Rater-8, Rater-10) pairing, for which overall agreement was 48.7%, and the *a priori* UniODA model achieved ESS=52.2 (*p*<0.0001), indicating relatively strong inter-observer agreement. Sparse data prevented jackknife analysis.

Third-strongest inter-observer reliability was obtained for the (Rater-8, Rater-9) pairing, for which overall agreement was 48.2%, and the *a priori* UniODA model achieved ESS=45.9 (*p*<0.0001), indicating moderate inter-observer agreement. Sparse data prevented jackknife analysis.

The fourth-strongest inter-observer reliability was obtained for the (Rater-6, Rater-9) pairing, for which overall agreement was 45.5%, and the *a priori* UniODA model achieved ESS=43.5 (*p*<0.0001), indicating moderate inter-observer agreement. Consistent performance was obtained in jackknife analysis.

Fifth-strongest inter-observer reliability was obtained for the (Rater-7, Rater-8) pairing, for which overall agreement was 51.3%, and the *a priori* UniODA model achieved ESS=40.4 (*p*<0.0001; moderate agreement). Overall agreement fell to 50.8%, and ESS fell to 27.9 (moderate agreement) in jackknife analysis.

Sixth-strongest inter-observer reliability was obtained for the (Rater-7, Rater-10) pairing: overall agreement=49.5%; ESS=34.0 (*p*<0.0001; moderate agreement). Overall agreement fell to 49.0%, and ESS fell to 21.5 (relatively weak agreement) in jackknife analysis.

The seventh-strongest inter-observer reliability was obtained for the (Rater-6, Rater-7) pairing, for which overall agreement was 42.0%, and the *a priori* UniODA model achieved ESS=31.1 (*p*<0.0001), indicating moderate inter-observer agreement. Consistent performance was obtained in jackknife analysis.

The eighth-strongest inter-observer reliability was obtained for the (Rater-9, Rater-10) pairing: overall agreement=49.0%; ESS= 29.9 (moderate agreement); *p*<0.0001. Overall agreement fell to 48.5%, and ESS fell to 26.8 (moderate agreement) in jackknife analysis.

Ninth-strongest inter-observer reliability was obtained for the (Rater-6, Rater-8) pairing: overall agreement=44.7%; ESS=25.4 (nearing the lower bound of moderate agreement); *p*< 0.0001. Consistent performance was obtained in jackknife analysis.

Finally, the weakest *a priori* model was obtained for the (Rater-6, Rater-10) pairing: overall agreement=37.0%; ESS=22.8 (relatively weak agreement); *p*<0.0001. Overall agreement fell to 36.5%, and ESS fell to 17.8 (relatively weak agreement) in jackknife analysis.

Table 2 summarizes the ESS achieved by the UniODA model for the training analyses involving the CTAS triage ratings.

Table 2: CTAS inter-observer ESS results

	Rater-7	Rater -8	Rater-9	Rater-10
Rater-6	31.1	25.4	43.5	22.8
Rater-7		40.4	53.6	34.0
Rater-8			48.2	52.2
Rater-9				29.9

Note: Tabled is ESS for training analysis predicting ratings of one rater given ratings of the other rater: ESS=0 is the level of agreement expected by chance; ESS=100 is perfect agreement.

Copyright 2014 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

As was discovered in the analysis of ESI-based triage codes, findings obtained in the analysis of CTAS-based triage codes clearly demonstrate that inter-observer agreement is far from perfect. Although two of the ten models identified relatively strong inter-observer agreement, another model was relatively weak, and two models failed the experimentwise criterion for statistical significance. However, in contrast to the exploratory/degenerate UniODA models required in analysis of ESI-based triage ratings, the *a priori* hypothesis was successfully tested for all ten CTAS-based rater pairings. It is concluded that prior kappa-based estimates indicating almost perfect inter-observer reliability of CTAS triage scores are untenable.

Comparing the Strength and Consistency of Inter-Observer Reliability of ESI and CTAS Triage Codes

The ESS values achieved by the inter-observer UniODA models were compared between the ESI (Table 1) and the CTAS (Table 2) using UniODA. Triage algorithm was treated as a binary class variable, and ESS as an ordered attribute (no *a priori* hypothesis was specified). The model achieved ESS=20.0 (a relatively weak effect), p>0.99, indicating that the ESS values of the inter-observer reliability models didn't discriminate triage algorithm. Thus, the inter-observer reliabilities achieved for the ESI and the CTAS were comparably mediocre.

The number of UniODA models which were consistent with the *a priori* hypothesis was compared between the ESI and the CTAS using UniODA. Triage algorithm was treated as the binary class variable, and whether or not the *a priori* model fit the triage data for the pair was treated as the binary attribute (no directional hypothesis was specified). The model achieved ESS=60.0 (relatively strong effect), p<0.011 (statistically significant at the generalized criterion, but *not significant* at the experimentwise criterion). The CTAS inter-observer models were thus significantly more consistent with the *a priori* hypothesis compared to ESI inter-observer models.

In Table 1 it is readily apparent that Rater-4 and in particular Rater-5 are using the ESI algorithm in a different manner than the other three raters. As seen in Figures 2-4, the effect of this difference is focused on the more serious cases having triage codes of 1 and 2. In light of the importance[29] of these particular codes in terms of real-time ED operational throughout and patient well-being, additional training in the identification and discrimination of code-1 and code-2 patients may be warranted. However, it is interesting that although all raters had prior experience using the CTAS to triage emergency patients, versus no prior experience using the ESI, no difference was found in level of inter-observer agreement (assessed as ESS) between algorithms. Instead the difference was manifest in terms of the number of models that supported the *a priori* hypothesis. The effect of insufficient experience using the ESI triage algorithm was therefore the observed omission, compression, polarization, and regression anomalies identified in the ESI ratings, but not found in the CTAS ratings.

References

[1]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC: APA Books.

[2]Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, *1*, 10-22.

[3]Yarnold PR (2013). Comparing attributes measured with "identical" Likert-type scales in single-case designs with UniODA. *Optimal Data Analysis, 2*, 148-153.

[4]Grimm LG, Yarnold PR (Eds.). *Reading and understanding multivariate statistics*. Washington, D.C.: APA Books, 1995.

[5]Grimm LG, Yarnold PR (Eds.). *Reading and understanding more multivariate statistics*. Washington, D.C.: APA Books, 2000.

[6]Yarnold PR (2014). UniODA *vs*. t-Test: Comparing two migraine treatments. *Optimal Data Analysis*, *3*, 6-8.

[7]Bryant FB, Yarnold PR (2014). Finding joy in the past, present, and future: The relationship between Type A behavior and savoring beliefs among college undergraduates. *Optimal Data Analysis*, *3*, 36-41.

[8]Yarnold PR (2010). UniODA *vs*. chi-square: Ordinal data sometimes feign categorical. *Optimal Data Analysis*, *1*, 62-65.

[9]Yarnold PR (2014). UniODA *vs*. chi-square: Audience effect on smile production in infants. *Optimal Data Analysis*, *3*, 3-5.

[10]Yarnold PR (2014). UniODA *vs*. chi-square: Discriminating inhibited and uninhibited infant profiles. *Optimal Data Analysis*, *3*, 9-11.

[11]Yarnold PR (2014). "Breaking-up" an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis*, *3*, 19.

[12]Yarnold PR (2010). Unconstrained covariates in CTA.*Optimal data Analysis, 1*, 38-40.

[13]Yarnold PR (2010). Aggregated *vs*. referenced categorical attributes in UniODA and CTA. *Optimal Data Analysis, 1*, 46-49.

[14]Reynolds HT (1977). *The analysis of cross-classifications*. New York: Free Press.

[15]Yarnold PR (2014). UniODA *vs*. kappa: Evaluating the long-term (27-year) test-retest reliability of the Type A Behavior Pattern. *Optimal Data Analysis*, *3*, 14-16.

[16]Yarnold PR (2014). UniODA *vs*. weighted kappa: Evaluating concordance of clinician and patient ratings of the patient's physical and mental health functioning. *Optimal Data Analysis*, *3*, 12-13.

[17]Feingold M (1992). The equivalence of Cohen's kappa and Pearson's chi-square statistics in the 2 x 2 table. *Educational and Psychological Measurement*, *52*, 57-61.

[18]Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JA (2004). Reliability and validity of scores on the Emergency Severity Index Version 3. *Academic Emergency Medicine*, *11*, 59-65.

[19]Worster A, Gilboy N, Fernandes CM, Eitel D, Eva K, Gleister R., Tanabe P (2004). Assessment of inter-observer reliability of two five-level triage and acuity scales: A randomized controlled trial. *Canadian Journal of Emergency Medicine*, *6*, 240-245.

[20]Strube MJ (2000). Reliability and generalizability theory. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding more multivariate statistics*. Washington, D.C.: APA Books, pp. 23-66.

[21]Grafstein E (2004). Close only counts in horseshoes and… triage? (Commentary). *Canadian Journal of Emergency Medicine*, *6*, 395-396.

[22]Fan J, Upadhye S, Woolfrey K (2006). ESI and CTAS (Letter). *Canadian Journal of Emergency Medicine*, *6*, 395-396.

[23]Bryant FB, Yarnold PR (1995). Principal components analysis and exploratory and confirmatory factor analysis. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding multivariate statistics*. Washington, D.C.: APA Books, pp. 99-136.

[24] The study was designed and data collected by my friend and colleague, Dr. David R. Eitel (deceased), an emergency medicine physician with a background in OR/MS—a strong proponent of optimal methodologies. One of the developers of the ESI, he was excited about this project. This research report represents only a portion of the Results section of the article that he would have produced, but in his absence I decided to deliver what I am able.

[25] Soltysik RC, Yarnold PR. (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, 2, 194-197.

[26] Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the wheat. *Optimal Data Analysis*, 2, 202-205.

[27] Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis*, 2, 220-221.

[28] Effects having $\text{ESS}<25$ are weak; $25\leq\text{ESS}<50$ are moderate; $50\leq\text{ESS}<75$ are relatively strong; and effects having $\text{ESS}\geq75$ are very strong.[1]

[29] Yarnold PR, Soltysik RC (2014). Emergency Severity Index (Version 3) score predicts hospital admission. *Optimal Data Analysis*, *3*, 20-22.

Author Notes

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., Suite 163
Chicago, IL 60646

eMail: Journal@OptimalDataAnalysis.com