Optimal Data Analysis
Vol. 3 (April 14, 2014), 50-54

2155-0182/10/$3.00

How to Assess the Inter-Method (Parallel-Forms) Reliability of Ratings Made on Ordinal Scales: 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-method 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. UniODA revealed moderate levels of inter-method agreement, which is theoretically consistent with recent findings of moderate inter-observer reliability for triage ratings derived using both algorithms.

The question of how best to statistically assess *inter-observer* or *inter-rater*[1] reliability recently arose in an investigation of triage codes made by two independent nurses utilizing the *identical algorithm*, vis-à-vis a scale offering five ordinal values as measurement options.[2]

The related issue addressed presently is how best to statistically assess *inter-method* or *parallel-forms*[1] reliability of assessments made by two nurses using *different algorithms* to assign triage scores using scales offering five ordinal measurement options. Conventional statistical methodologies are ineffectual for such ordinal data: nominal methods such as chi-square ignore ordinal information in the data[3-5] and parametric methods such as analysis of variance were developed for use with interval data, and assume sample data are distributed in a specific manner and fulfill other assumed requirements.[6,7]

In contrast to implacably challenged suboptimal methods, UniODA is ideally-suited to conduct exact, optimal (maximum-accuracy) statistical analysis for data measured on ordinal scales.[8-11] The use of UniODA to assess inter-observer, test-retest and parallel-forms reliability, and in structural decomposition of ordinal reliability data (analogous to principal components analysis except that accuracy, instead of variance[12], is explicitly maximized), is well-

2155-0182/10/$3.00

documented.[8,13,14] In the present context UniODA tests the *a priori* hypothesis that triage codes (integers ranging from 1 to 5)—which are independently assigned to patients by two nurses using different algorithms—are consistent. This *a priori* hypothesis is defined in UniODA (using the *directional* command[8]) 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

Described[15] 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 were trained in use of the 5-point *E*mergency *S*everity *I*ndex (ESI) Version 3 triage algorithm. The other five nurses 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 "…200 case scenarios abstracted from prospectively collected local ED cases" (p. 242).

Data evaluated in the present study were analyzed previously[15] using generalizability theory[16]: "The two triage scales appear to be in moderate agreement with one another, as indicated by an inter-test generalizability of 0.58" (p. 243). These data[17] are reanalyzed presently using UniODA[8] run on MegaODA software.[18-20]

Results and Discussion

For each unique rater pair involving different algorithms, UniODA evaluated the *a priori* hypothesis that the triage codes agreed (Figure 1). This model fit the data of 13 of the 25 unique rater pairings (Table 1).

Table 1: Rater pairings consistent with the *a priori* hypothesis

ESI Rater	CTAS Rater	Training Agreement	ESS	Jackknife Agreement	ESS
1	1	34.5%	35.0		
	2	49.0%	46.7		
	3	41.2%	27.3	40.7%	14.8
	4	48.5%	46.5		
	5	48.5%	33.8	48.0%	21.3
2	1	34.0%	32.2		
	2	53.0%	47.2		
	3	42.7%	25.5	42.2%	17.1
	4	49.0%	39.6		
	5	53.5%	36.0	53.0%	27.7
3	1	39.0%	41.9	NA	
	2	53.0%	49.8	NA	
	4	50.0%	46.5	NA	

Note: Jackknife agreement was stable unless specified; NA= not available due to sparse data for one or more triage codes. All p were statistically significant at the experimentwise criterion[8] (all $p<0.0001$). ESS values given in red indicate relatively weak association, otherwise the ESS values indicate moderate association.[8]

For all 12 remaining pairings 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.[8]

2155-0182/10/$3.00

For seven pairings the exploratory UniODA model which is illustrated in Figure 2 was identified.

Figure 2: Exploratory UniODA Model

Rater-A		Rater-B
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[8] (specifically, Type A; pp. 137-138). Triage codes in these pairings demonstrate *local regression* at lower levels of the scale (over the two most severe triage codes), but are stable over the remaining range of the scale. Table 2 summarizes the findings of these analyses.

Table 2: Rater pairings consistent with the exploratory UniODA model in Figure 2

ESI Rater	CTAS Rater	Training Agreement	ESS	$p <$
3	3	38.2%	33.5	0.0008
	5	33.5%	28.5	0.04*
4	1	24.1%	33.2	0.007
	2	33.2%	35.7	0.0008
	3	32.8%	33.6	0.003
	4	41.7%	40.4	0.0001
	5	35.2%	34.6	0.002

Note: Jackknife analysis was not possible for these models due to sparse data for one or more triage codes. The asterisk indicates *p* was statistically significant at the generalized (per-comparison) criterion, but *wasn't* statistically significant at the experimentwise criterion.[8] The ESS values indicate moderate association.[8]

For the remaining five rater pairs 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.[8] The model illustrated in Figure 3 emerged for all five pairings.

Figure 3: Exploratory Degenerate UniODA Model

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

In Figure 3 the UniODA model reveals collapsing (*local compression*) for the most serious cases (triage codes<3), but consistent assignments for triage codes $\geq$3. The model is degenerate because no predictions of triage code 2 are made for Rater-B. Table 3 summarizes the findings of these analyses.

Table 3: Rater pairings consistent with the exploratory degenerate UniODA model in Figure 3

ESI Rater	CTAS Rater	Training Agreement	ESS	Jackknife Agreement	ESS
5	1	33.5%	23.6		
	2	46.5%	32.6		
	3	36.7%	25.0		
	4	46.5%	31.5		
	5	49.5%	35.4	43.5%	21.6

Note: Jackknife agreement was stable unless specified. All *p* were statistically significant at the experimentwise criterion[8] (all $p<0.002$). ESS values given in red indicate relatively weak association, and otherwise the ESS values indicate moderate association.[8]

2155-0182/10/$3.00

The findings indicate moderate inter-method agreement for ESI- and CTAS-based triage codes. No rater pair achieved relatively strong agreement, one pair returned relatively weak agreement, and 24 pairs evidenced moderate levels of agreement. The *a priori* UniODA model was feasible for 13 rater pairs, but 12 pairs revealed patterns of *consistent disagreement*. Five of the exploratory models identified compression inconsistencies, and seven models identified local regression for the most serious emergency medicine cases. The finding for one rater pair wasn't statistically significant at the experimentwise criterion.

As expected[1], the weakest training ESS obtained for inter-method analyses (23.6) was comparable to the weakest training ESS values obtained in prior research[2] for inter-observer analyses (22.8 for CTAS ratings; 28.6 for ESI ratings). And, also as expected[1], the strongest ESS observed for inter-method UniODA models (49.8) was lower than strongest ESS observed for inter-observer models for ESI (59.9) or CTAS (53.6) triage ratings.

References

[1]Magnusson D (1966). *Test theory*. Reading, MA: Addison-Wesley.

[2]Yarnold PR (2014). 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. *Optimal Data Analysis*, *3*, 42-49.

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

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

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

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

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

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

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

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

[11]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.

[12]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.

[13]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.

[14]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.

2155-0182/10/$3.00

[15]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.

[16]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.

[17]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 elected to produce what I am able.

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

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

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

Author Notes

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

eMail: Journal@OptimalDataAnalysis.com