Optimal Data Analysis
Vol. 5 (November 7, 2016), 179-183

2155-0182/10/$3.00

Novometric *vs*. ODA Reliability Analysis *vs*. Polychoric Correlation with Relaxed Distributional Assumptions: Inter-Rater Reliability of Independent Ratings of Plant Health

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

Inter-rater reliability of independent ordinal ratings of plant health is assessed by various methods, and compared.

Data in Table 1 summarize health ratings of 460 trees and shrubs made by a pair of independent raters using an ordinal scale on which increasing integers indicate increasingly healthy plants (1= least healthy; 6=most healthy).[1,2]

Table 1: Plant Health Ratings by Two Judges

	Rater A					
Rater B	1	2	3	4	5	6
1	30	1	0	0	0	0
2	0	10	2	0	0	0
3	0	4	8	3	1	0
4	0	3	3	37	9	0
5	0	0	1	25	71	49
6	0	0	0	2	20	181

Original analysis estimated the polychoric correlation to be 0.95, but the statistically significant G^2 model "badness-of-fit" statistic of 57.33 (df=24, p<0.001) called the assumptions underlying the standard polychoric correlation model into question.[1] Accordingly, "…the assumption of a normally distributed latent trait was relaxed, and a latent trait model with a non-parametric latent trait distribution was fit to the data. The distribution was represented as six equally-spaced locations (located latent classes) along a unidimensional continuum, and the density at each location (latent class prevalence) was estimated. Model fit, assessed using the G^2 statistic was 15.65 on 19 df (p = 0.66). The correlation of each variable with the latent trait was 0.963. This value squared, 0.927, estimates the correlation of the raters if ratings were made using a continuous scale. This generalization of the polychoric correlation…is called the latent correlation between the raters."[1]

The estimated latent correlation may be used to predict Rater A's ratings using B's ratings (and vice versa), and to compute corresponding predictive accuracy normed *vs*. chance (ESS) yielded by the latent correlation model.[3-7] The standardized regression model is $z_y^*=rz_x$, or $z_y^*=0.927z_x$, where z_y^* is the predicted z_y score for an observation. This model yielded the same

2155-0182/10/$3.00

performance achieved by the confirmatory ODA reliability model discussed ahead.

ODA-Based Reliability Analysis

The use of ODA to evaluate inter-rater and inter-device reliability is discussed in detail elsewhere.[7-11] In this approach the first *a priori* hypothesis tested is that ratings made by a pair of raters agree: that is, that the data fall into the major diagonal running from the upper left-hand to lower right-hand corner in Table 1. Thus the confirmatory hypothesis correctly predicts ratings in the major diagonal in Table 2: NA=the

Table 2: Confusion Table for Initial ODA Test of Confirmatory Hypothesis that Raters' Ratings Agree

		Predicted Rater A							
		1	2	3	4	5	6	NA	Sens
Actual A	1	30	**1**	0	0	0	0	31	96.77%
	2	0	10	**2**	0	0	0	12	83.33%
	3	0	4	8	3	1	0	16	50.00%
	4	0	3	3	37	9	0	52	71.15%
	5	0	0	1	25	71	49	146	48.63%
	6	0	0	0	2	20	181	203	89.16%
	NP	30	18	14	67	101	230		ESS = 67.81
	PV	100.00%	55.56%	57.14%	55.22%	70.30%	78.70%		p < 0.001
									D = 2.85

		Predicted Rater B							
		1	2	3	4	5	6	NA	Sens
Actual B	1	30	0	0	0	0	0	30	100.00%
	2	**1**	10	4	3	0	0	18	55.56%
	3	0	**2**	8	3	1	0	14	57.14%
	4	0	0	3	37	25	2	67	55.22%
	5	0	0	1	9	71	20	101	70.30%
	6	0	0	0	0	49	181	230	78.70%
	NP	31	12	16	52	146	203		ESS = 63.38
	PV	96.77%	83.33%	50.00%	71.15%	48.63%	89.16%		p < 0.001
									D = 3.47

2155-0182/10/$3.00

number of actual observations; PP=number of predicted observations; Sens=model sensitivity; PV=model predictive value.

If the initial confirmatory model is found to be statistically reliable, in the next (second) step of ODA reliability analysis the correctly classified observations in step one are deleted (corresponding cell values in the major diagonal of Table 1 are set to zero), and an exploratory ODA analysis is run in an effort to identify any statistically reliable bias that may exist in the remaining residual rating data.[7-11] Red font is used in Table 2 to indicate values that must be modeled in step two to prevent a degenerate model omitting one or more class categories.[7,8] The exploratory model predicting Rater A responses in step two will be degenerate due the absence of any ratings of "1" made by Rater A.

When the second step of the reliability analysis—in which exploratory ODA analysis is applied to the step one residual (misclassified) data—was applied to predict residuals for Rater A, LOO analysis was not possible (at least two observations per class category are required[7,8]), and no statistically reliable model emerged. This finding suggests there is no reliable pattern of bias underlying the residuals for Rater A after removing inter-rater agreement with Rater B.

For the second step of the reliability analysis attempting to predict Rater B residuals using A's responses, the following model was obtained (ODA models are not symmetric[12,13]):

Rater's A Rating	Rater's B Predicted Rating
<3	3
3	4
4	2
5	6
6	5

The result of using this model to classify residuals is shown in Table 3 for both training and LOO analysis. As seen, the small sample (and associated weak statistical power[7]), and the resulting restricted (indicated using red) values, are problematic for further decomposition. Furthermore, if a third step was to be conducted then there would be no actual ratings of 6 by Rater B, and thus a degenerate model would be necessitated.

As seen, although predictive accuracy (ESS) declined to a relatively weak level in

Table 3: Confusion Table for Exploratory ODA Training (Top) and LOO (Bottom) Analysis Predicting Rater B's Step One Residuals

		Predicted Rater B (*Training Analysis*)						
		2	3	4	5	6	NA	Sens
Actual Rater B	2	3	**1**	4	0	0	8	37.50%
	3	3	2	0	0	1	6	33.33%
	4	0	0	3	2	25	30	10.00%
	5	**9**	0	1	20	0	30	66.67%
	6	0	0	0	0	49	49	100.00%
	NP	15	3	8	22	75		ESS = 36.87
	PV	20.00%	66.67%	37.50%	90.91%	65.33%		$p < 0.001$
								D = 8.56

2155-0182/10/$3.00

		Predicted Rater B (*LOO Analysis*)						
		2	3	4	5	6	NA	Sens
	2	0	4	4	0	0	8	0.00%
A								
c	3	5	0	0	0	1	6	0.00%
t								
u	4	0	0	3	2	25	30	10.00%
a								
l	5	9	0	1	20	0	30	66.67%
B	6	0	0	0	0	49	49	100.00%
	NP	14	4	8	22	75		ESS = 19.17
	PV	0.00%	0.00%	37.50%	90.91%	65.33%		p < 0.001
								D = 21.08

LOO cross-generalizability analysis, the model sensitivity was perfect for the most extreme positive health rating of 6, and was strong for ratings of 5 that were misclassified by the confirmatory model.

While the original legacy analysis and ODA confirmatory model were identical—and the model was most accurate in predicting the lowest ratings of plant health, for Rater B the ODA exploratory analysis further improves classification of highest ratings of plant health. Integrating the LOO confusion tables for the confirmatory and exploratory models for Rater B yields ESS=77.08, p<0.001, D=2.97 (for 10 model strata). Whether the reliable bias detected in exploratory LOO analysis for Rater B can be extinguished vis-à-vis increased training, or if such bias will cross-generalize to comparisons involving other raters, are questions that cannot be answered statistically using the present data.

Novometric Reliability Analysis

The use of novometric analysis to evaluate the association between ordered class variables and ordered attributes (e.g., A's and B's ratings) is discussed elsewhere.[7,14,15] However, this is the first demonstration of novometric evaluation of inter-rater reliability.

The first analysis modeled Rater A's ratings (class variable) as a novometric function of B's ratings (attribute). The globally optimal (GO) model for this application was: if Rater B's rating=1, predict Rater A's rating=1; otherwise predict Rater A's rating>1. Stable in LOO analysis, this model correctly classified 30/31 (96.8%) of Rater A's ratings of 1, and 429/429 (100%) of Rater A's ratings greater than 1: very strong ESS=96.8, p<0.001, D=0.067.

The second analysis modeled B's ratings (class variable) as a novometric function of A's ratings (attribute). The globally optimal (GO) model for this application was: if Rater A's rating=1, predict Rater B's rating=1; otherwise predict Rater B's rating>1. Stable in LOO analysis, this model correctly classified 30/30 (100%) of Rater B's ratings of 1, and 429/430 (99.8%) of Rater B's ratings greater than 1: very strong ESS=99.8, p<0.001, D=0.0046.

Not unexpectedly the novometric models were most accurate and straightforward, and both models revealed the underlying metric shared by this pair of raters was binary. That is, both GO models indicated that a more granular measurement scale than was identified presently (a rating of 1 *vs.* greater than 1) yields suboptimal predictive accuracy. Expressed in applied terms this means the independent raters agreed

2155-0182/10/$3.00

almost perfectly whether a plant had the lowest possible level of health, or not—and any effort to further granulate the model in this application only decreases normed predictive accuracy.

References

[1]Uebersax JS (2006; accessed 11/5/2016). The tetrachoric and polychoric correlation coefficients. *Statistical Methods for Rater Agreement*. http://john-uebersax.com/stat/tetra.htm

[2]Hutchinson TP (2000). Assessing the health of plants: Simulation helps us understand observer disagreements. *Environmetrics, 11*, 305-314.

[3]Yarnold PR, Bryant FB, Soltysik RC (2013). Maximizing the accuracy of multiple regression models via UniODA: Regression *away from* the mean. *Optimal Data Analysis, 2*, 19-25.

[4]Yarnold PR (2013). Maximum-accuracy multiple regression analysis: Influence of registration on overall satisfaction ratings of emergency room patients. *Optimal Data Analysis*, *2*, 72-75.

[5]Yarnold PR (2013). Assessing technician, nurse, and doctor ratings as predictors of overall satisfaction ratings of Emergency Room patients: A maximum-accuracy multiple regression analysis. *Optimal Data Analysis*, 2, 76-85.

[6]Yarnold PR (2015). Maximizing ESS of regression models in applications with dependent measures with domains exceeding ten values. *Optimal Data Analysis*, *4*, 12-13.

[7]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

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

[9]Yarnold, P.R. (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.

[10]Yarnold, P.R. (2014). How to assess the inter-method (parallel-forms) 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*, 50-54.

[11]Yarnold. P.R. (2015). Estimating inter-rater reliability using pooled data induces paradoxical confounding: An example involving Emergency Severity Index triage ratings. *Optimal Data Analysis*, *4*, 21-23.

[12]Yarnold PR (2016). Matrix display of pairwise novometric associations for ordered variables. *Optimal Data Analysis, 5*, 94-102.

[13]Yarnold PR, Batra M (2016). Matrix display of pairwise novometric associations for mixed-metric variables. *Optimal Data Analysis*, *5*, 104-107.

[14]Yarnold PR, Linden A (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis. *Optimal Data Analysis, 5*, 65-73.

[15]Yarnold PR, Bennett CL (2016). Novometrics *vs*. correlation: Age and clinical measures of PCP survivors. *Optimal Data Analysis, 5*, 74-78.

Author Notes

This study analyzed publically available data. No conflict of interest was reported.