Optimal Data Analysis
Vol. 8 (May 23, 2019), 107-111

2155-0182/10/$3.00

Regression *vs.* Novometric-Based Assessment of Inter-Examiner Reliability

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

Four examiners independently recorded the DMFS (decayed, missing, filled surfaces) scores of ten patients.[1] Inter-examiner correspondence of DMFS scores was evaluated using Pearson correlation and novometric analysis. Whereas essentially perfect correlation models were unable to accurately predict DMFS scores in training analysis, novometric models were consistently perfect in both training and reproducibility analysis.

DMFS scores were given by four independent examiners to a sample of ten patients. Analysis of variance found that rater C assigned a higher mean rating for patients than the other raters.[1] Pairwise analysis conducted by relative optimal threshold ODA found the distribution of rater C's ratings was virtually perfectly greater than corresponding (non-discriminable) ratings made by other raters.[2] Both findings reveal differential measurement bias exists because rater C differs systematically from the other raters.[1,3]

Accordingly, a statistical issue requiring attention is circumventing Simpson's paradox—which in a moderate case can over- or underestimate effect strength, and in a severe case can miss a true effect or identify a false effect.[4-7] In the present design paradoxical confounding may arise due to a difference in the rating means[1] or response distributions[2] between the raters, or if there isn't linear correspondence between raters' ratings for the sample.[8] This paper assesses the latter issue, examining inter-rater reliability of raters' ratings by evaluating how ratings made by independent raters for each of ten patients (Table 1) are related to each other.[9]

Table 1: Inter-Rater Reliability Study Data: DMFS Scores of Four Raters for Ten Patients[1]

	Rater			
Patient	A	B	C	D
1	8	7	11	7
2	13	11	15	13
3	0	0	2	1
4	3	6	9	6
5	13	13	17	10
6	19	23	27	18
7	0	0	1	0
8	2	0	4	5
9	18	20	22	16
10	5	3	8	3

Assessing Inter-Examiner Agreement: Pearson Correlation (Regression) Analysis

Figure 1 shows raw DMFS ratings of all four raters (A=**Blue**; B=**Red**; C=**Green**; D=**Purple**) for all ten patients. Visual examination suggests raters agreed strongly about DMFS scores of the patients (inter-rater scores don't vary substantially for most patients). Patients #6 and #9 had greatest DMFS scores, and patients #3 and #7

2155-0182/10/$3.00

had lowest scores. The inequality dominance of scores for rater C (green) is clearly seen, but the magnitude of differences (percent difference in DMFS score) is comparatively modest.

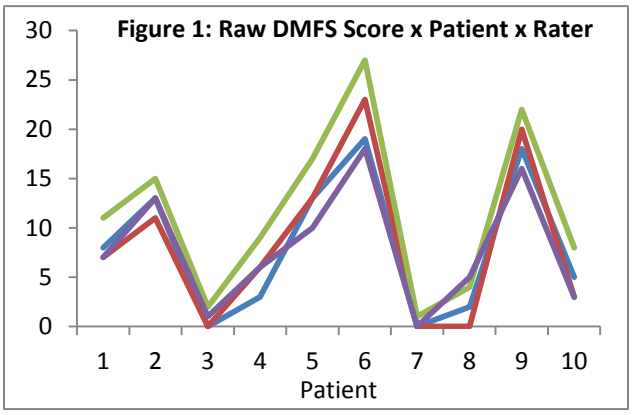

Figure 2 is the corresponding illustration for ipsative standard scores.[10,11] Note how dispersion in DMFS scores seen between raters in raw data (Figure 1) is reduced by use of ipsative standardization (Figure 2): only data for patients #2, #4 and #8 remain modestly variable. Note also that the near perfect inequality dominance of rater C *vs*. the other raters observed for raw DMFS scores[2] vanishes in the ipsatized data.

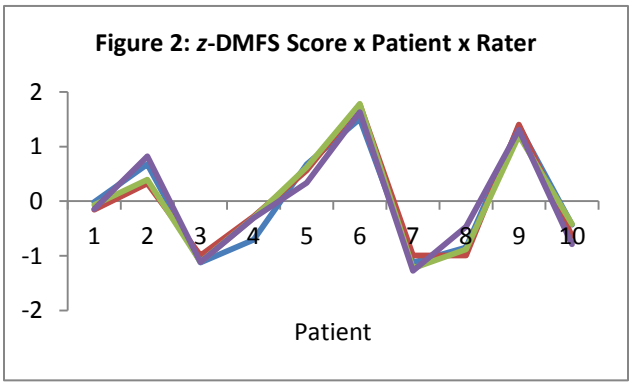

Table 2 presents the Pearson correlation coefficients between ratings of the ten patients for different pairs of raters: the estimated inter-rater reliability coefficients approached perfect, ranging between r=0.96 (R^2=92.2) and r=0.99 (R^2=98.0; all p<0.0001). Identical correlations were obtained for raw and ipsatized data.

Table 2: Inter-Rater Reliability (Pearson *r*) Across Ten Patients for Pairs of Raters

Rater	B	C	D
A	0.9718	0.9796	0.9651
B		0.9897	0.9598
C			0.9656

Accuracy of the strongest model (B and C) was assessed. Given the near perfect *r* for these ratings, regression models were expected to produce extremely accurate predictions of DMFS scores. Ipsative *z*-scores were expected to yield the most accurate predictions. Findings are given in Table 3: Pt=patient; Pred=rating predicted by the regression model; Real=actual rating; and Error=(Pred-Real)/Real x 100% (three missing error values occur in rater B's raw data results due to division by zero).

Table 3: Predicting Rater Rating by Regression

		Raw Data			*Ipsative Data*		
Pt	Rater	Pred	Real	Error	Pred	Real	Error
1	B	7.72	7	10.3	-0.07	-0.16	-56.2
1	C	10.27	11	-6.6	-0.15	-0.07	114.3
2	B	11.55	11	5.0	0.39	0.32	21.9
2	C	14.37	15	-4.2	0.32	0.39	-17.9
3	B	-0.88	0	----	-1.10	-1.00	10.0
3	C	3.10	2	55.0	-0.99	-1.11	-10.8
4	B	5.81	6	-3.2	-0.30	-0.28	7.1
4	C	9.24	9	2.7	-0.27	-0.30	-10.0
5	B	13.46	13	3.5	0.62	0.56	10.7
5	C	16.41	17	-3.4	0.56	0.63	-11.1
6	B	23.02	23	0.1	1.77	1.77	0
6	C	26.66	27	-1.3	1.75	1.79	-2.2
7	B	-1.83	0	----	-1.22	-1.00	22.0
7	C	3.10	1	210.0	-0.99	-1.23	-19.5
8	B	1.03	0	----	-0.87	-1.00	-13.0
8	C	3.10	4	-22.5	-0.99	-0.88	12.5
9	B	18.25	20	-8.8	1.19	1.41	-15.6
9	C	23.58	22	7.2	1.39	1.21	14.9
10	B	4.86	3	62.0	-0.41	-0.64	-35.9
10	C	6.18	8	-22.8	-0.63	-0.42	50.0

2155-0182/10/$3.00

As was expected, the use of ipsative z-scores reduced overall absolute mean percent error (±22.8% for 20 models) compared to raw rating data (±25.2% for 17 models). Nevertheless, these values translate to prediction bands 45.5% and 50.4% wide, respectively. Eight models using raw scores and two using ipsative scores have training performance falling within a 10% band of accurate prediction; three of both model types have performance falling within a 20% band; and one raw score and six ipsative score models have performance falling within a 30% band. These results fail to satisfy the anticipated "extreme accuracy" criterion.

Novometric Analysis

Table 4 presents the novometric models which emerged between ratings of the ten patients for different pairs of raters.[12] All reliability models yielded perfect accuracy in training and one-sample jackknife leave-one-out (LOO) analysis (exact *p*'s<0.00397). All models classified half of each rater's responses above and half below optimal cutpoints. Identical values of ESS and *p* emerged when analyzing the raw and ipsatively standardized data.

Table 4: Novometric Models Relating Ratings: For All Models, Jackknife ESS=100, $p<0.00397$

	Rater A	Rater B	Rater C
Rater B	If B≤6 then A≤5 If B>6 then A>5		
Rater C	If C≤9 then A≤5 If C>9 then A>5	If C≤9 then B≤6 If C>9 then B>6	
Rater D	If D≤6 then A≤5 If D>6 then A>5	If D≤6 then B≤6 If D>6 then B>6	If D≤6 then C≤9 If D>6 then C>9

These optimal inter-examiner reliability models are symmetric about the major diagonal. For example, to predict responses of rater A (as class variable) based on the responses of rater B (as attribute), the model is: if B≤6 predict A≤5, otherwise predict A>5. To predict responses of rater B (class variable) based on responses of rater A (attribute), the model is: if A≤5 predict B≤6, otherwise predict B>6.

Exercises described herein are useful in the exposition of computational and interpretive aspects of optimal inter-rater reliability analysis. However, present data clearly are inadequate if considered from a production standpoint, since the first of four axioms of novometric statistical theory is that the sample N provides sufficient (desired) statistical power to test the *alternative hypothesis* (90% power is a minimum standard in research presently).[13-16]

Nevertheless, some qualitative aspects of results in Table 4 are of interest. For example, the optimal class and attribute thresholds used in models involving raters A, B and D all differ by only a single DMFS point—and by zero points for raters B and D. In contrast, models with rater C had optimal class and attribute thresholds which differed by three or more DMFS points: rater C always had the threshold of greater value for raw scores, but not for ipsative *z* scores.

All optimal solutions which classified patient ratings perfectly in LOO validity analysis involved the use of single threshold values for both class and attribute. The models yielded perfect LOO-stable solutions due to synergy of small sample and high inter-patient variance in DMFS rating—making near-maximum-strength (perfect for novometrics, almost perfect for regression) inter-examiner correspondence achievable. Replicating this research with larger patient samples may identify multiple qualitative patient groups—such as high, moderate and low DMFS-score groups.[14]

All six novometric analyses involving different pairs of raters identified more than one optimal model which achieved 100% accurate classification and was stable in LOO analysis. Of the multiple optimal models identified, all six analyses identified a model which separated raters' ratings into balanced higher (N=5) and lower (N=5) classes (Table 4) and was selected *a priori* on the basis of having greatest statistical power.[16] For example, five LOO-stable models having perfect ESS were identified to predict responses of rater C based on responses of rater

2155-0182/10/$3.00

D (and *vice versa*): (1) if D=0 then C=0, otherwise C>0 (n_0=2, $n_{>0}$=8; p<0.023); (2) if D≤5 then C≤8, otherwise C>8 ($n_{\leq 5}$=4, $n_{>5}$=6; p< 0.0048); (3) if D≤6 then C≤9, otherwise C>9 ($n_{\leq 9}$=5, $n_{>9}$=5; p<0.0040); (4) if D≤7 then C≤11, otherwise C>11 ($n_{\leq 11}$=6, $n_{>11}$= 4; p<0.0052); and (5) if D≤13 then C≤17, otherwise C>17 ($n_{\leq 13}$=8, $n_{>13}$=2; p<0.023). These five perfectly accurate models identify DMFS scores used by raters C and D to identify patient groups having worsening increasing dental issues (moving left to right, model 1 to 5), or diminishing dental issues (moving right to left, model 5 to 1).

Perhaps the most surprising findings in this study are (a) that raters' DMFS ratings were correlated nearly perfectly, and (b) even so the regression models were unable to make accurate predictions in training analysis—netting 22.8% (ipsative *z* scores) to 25.2% (raw scores) overall absolute mean percent error in predicting DMFS scores. In contrast, for each pair of raters, novometric analysis found multiple models yielding 100% agreement in training and LOO analysis. An accumulating mass of new research shows regression analysis can and does make accurate predictions in random data[17,18] but has trouble in accurately predicting data for which strong[19-21] and even virtually perfect linear effects exist—as is the case presently.

References

[1]Fleiss JL (1986). *The design and analysis of clinical experiments*. New York, NY: Wiley (pp. 19-26).

[2]Yarnold PR (2019). Fixed *vs*. relative optimal discriminant thresholds: Pairwise comparisons of raters' ratings for a sample. *Optimal Data Analysis*, *8*, 103-106.

[3]Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice, 22,* 848-854. DOI: 10.1111/jep.12538

[4]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement, 56,* 430-442.

[5]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: Guidebook with software for Windows*. Washington, D.C.: APA Books.

[6]Yarnold PR (2013). Ascertaining an individual patient's *symptom dominance hierarchy*: Analysis of raw longitudinal data induces Simpson's Paradox. *Optimal Data Analysis*, *2*, 159-171.

[7]Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, *1*, 67-100.

[8]Yarnold PR (2019). When to evaluate a nonlinear model. *Optimal Data Analysis, 8,* 15-20.

[9]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.

[10]Yarnold PR, Soltysik RC (2013). Ipsative transformations are *essential* in the analysis of serial data. *Optimal Data Analysis*, *2*, 94-97.

[11]Mueser KT, Yarnold PR, Foy DW (1991). Statistical analysis for single-case designs: Evaluating outcomes of imaginal exposure treatment of chronic PTSD. *Behavior Modification, 15,* 134-155.

[12]Yarnold PR (2016). Matrix display of pairwise novometric associations for ordered variables. *Optimal Data Analysis, 5,* 94-101.

[13]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis, 6,* 26-42.

2155-0182/10/$3.00

[14]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, *3*, 55-77.

[15]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis*, *3*, 78-84.

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

[17]Linden A, Yarnold PR (2019). Some machine learning algorithms find relationships between variables when none exist -- CTA doesn't. *Optimal Data Analysis*, *8*, 64-67.

[18]Linden A, Yarnold PR (2019). Effect of sample size on discovery of relationships in random data by classification algorithms. *Optimal Data Analysis*, *8*, 76-80.

[19]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.

[20]Yarnold PR (2019). Regression *vs*. novometric analysis predicting income based on education. *Optimal Data Analysis*, *8*, 81-83.

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

Author Notes

No conflict of interest was reported.