Optimal Data Analysis
Vol. 2, Release 2 (November 1, 2013), 148-153

2155-0182/10/$3.00

# Comparing Attributes Measured with "Identical" Likert-Type Scales in Single-Case Designs with UniODA

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

Often at the advice of their physician, patients managing chronic disease such as fibromyalgia or arthritis, or those undergoing therapy in rehabilitation medicine or oncology, will record weekly or daily—sometimes even real-time ratings—of physical (e.g., pain, fatigue) and emotional (e.g., depression, anxiety) symptoms. Similarly, often at the advice of their coaches, athletes ranging from elite professionals to everyday people engrossed in an astonishing variety of group and customized personal fitness programs, record ratings of physical (e.g., vigor, focus) and emotional (e.g., anger, fear) states at the beginning and/or the end of workout sessions. In these types of applications, within a given study (also typically in most related studies in any given discipline, and across different disciplines), physical and emotional symptoms and states are all assessed using the same type of measuring scale, specifically an ordered Likert-type scale with 3-11 response categories. This paper shows how to employ UniODA to compare such symptom or state ratings to identify the *strongest and weakest symptoms/states* using data obtained across multiple measurements for an individual.

Data for this study were abstracted with permission from a computer log containing 297 sequential entries of an anonymous patient with fibromyalgia (FM), who voluntarily used an intelligent health diary in the prospective clinical trial of a web-based, self-monitoring and symptom management system known commercially as *SMARTLog*.[1] Patients in the trial could use SMARTLog daily if they wished, and though no users made daily entries over protracted periods, overall usage of and satisfaction with the system were high.[1,2] On every session the users rated symptoms often reported in the literature for FM patients, including pain and fatigue—two of the most prevalent FM symptoms. Symptoms were rated on an 11-point Likert-Type scale ranging from 0 ("not at all bothersome") to 10 ("extremely bothersome").

Analysis is conducted to determine if the patient rated these two symptoms similarly: that is, if these two symptoms were experienced with comparable intensity, or if instead the ratings of

2155-0182/10/$3.00

one symptom exceed ratings of the other symptom, and the former is described as being *dominant* or *primary* versus the latter.

Analyses comparing ratings on pain and fatigue are first conducted using raw data, and a second time using data ipsatively standardized into *z*-scores using mean and standard deviation (SD) computed for the individual's data.[3]

## Analysis of Raw Scores

Descriptive statistics for *pain*: mean= 4.60; SD=1.52; median=5; skewness=0.20; kurtosis=-0.53; coefficient of variation (CV)=33.2. For *fatigue*: mean=6.38; SD=1.44; median=6; skewness=-0.03; kurtosis= -0.59; and CV=22.6. Table 1 also provides the pain and fatigue rating distributions for this patient.

Table 1: Distributions of the Patient's Raw Pain and Fatigue Ratings

| Response Category | Pain | Fatigue |
|---|---|---|
| 2 | 26 | |
| 3 | 48 | 5 |
| 4 | 74 | 25 |
| 5 | 65 | 50 |
| 6 | 49 | 85 |
| 7 | 27 | 60 |
| 8 | 7 | 49 |
| 9 | 1 | 23 |

Note: Available response categories of 0, 1 and 10 *weren't* used by patient.

Construct a new data set (*new.txt*) with 2*n* observations (2*297=594), each of which forms a row in new.txt. First, copy all 297 *pain* ratings to new.txt (which at this point has 297 rows). Next, beneath these, copy all 297 *fatigue* ratings (new.txt now has 594 rows). Add the number "0" (an arbitrary class category dummy-code) delimited by a space to the beginning of *each* of the *first* (top) set of 297 rows, and then likewise add the number "1" (arbitrary dummy-code) delimited by a space to the beginning of *each* of the *second* (bottom) set of 297 rows.

Using new.txt as input, the *post hoc* hypothesis that the raw pain and fatigue measures were rated differentially by the patient is tested running the following UniODA[4] code (control commands indicated in red):

```
OPEN new.txt;
OUTPUT example.out;
VARS class rating;
CLASS class;
ATTR rating;
MCARLO ITER 100000;
LOO;
GO;
```

One hundred thousand Monte Carlo experiments were used to estimate *p*, the default setting in the ODA laboratory for exploratory data analysis involving UniODA with a binary class variable.[5] The model had strong statistical support, with nearly perfect confidence for target $p<0.001$ (estimated $p<0.00001$). Analysis was completed in 16 CPU seconds by UniODA[4] run on a 3 GHz Intel Pentium D microcomputer.

The UniODA model was: if rating is $\leq 5$ then predict the attribute is *pain*; otherwise predict *fatigue*. This model indicates fatigue ratings were greater than pain ratings. As seen in Table 1, 26+48+74+65=213 (71.7%) of 297 pain ratings were correctly predicted, as were 85+60+49+23=217 (73.1%) of 297 fatigue ratings. The confusion table for the model is seen in Table 2: overall, ESS=44.78 and ESP=44.79, reflecting moderate strength.[4] Model classification performance was consistent in LOO analysis.[4]

*All possible aggregated confusion tables* (ACT) were examined to assess if increasing the reliability of the ratings yields superior symptom discrimination.[6]

2155-0182/10/$3.00

Table 2: Confusion Table for UniODA Model Discriminating Pain and Fatigue Ratings

| | | Predicted Symptom | | |
|---|---|---|---|---|
| | | Pain | Fatigue | |
| Actual | Pain | 213 | 84 | 71.7% |
| Symptom | Fatigue | 80 | 217 | 73.1% |
| | | 72.7% | 72.1% | |

The first ACT analysis dropped the two middle response categories of 5 and 6 (see Table 1), yielding the ACT provided in Table 3 (58% of the total sample). Resultant ESS (62.4) and ESP (62.2) were both 39% greater than achieved in full-range analysis, and both in the domain of a strong effect.[4]

Table 3: First ACT Omitting Ratings of 5 and 6

| | | Predicted Symptom | | |
|---|---|---|---|---|
| | | Pain | Fatigue | |
| Actual | Pain | 148 | 35 | 80.9% |
| Symptom | Fatigue | 30 | 132 | 81.5% |
| | | 83.2% | 79.0% | |

The second ACT additionally dropped response categories of 4 and 7 yielding the ACT in Table 4: the ESS (83.8) and ESP (83.7) were very strong effects, 87% greater than full-range analysis and 35% greater than the first ACT, but only 27% of the total sample is classified—half of the number of the first ACT. The third and final ACT (Table 5) only used the most extreme ratings, 2 and 9. Resulting ESS (96.3) and ESP (95.8) were nearly perfect but only a meager 8% of the sample was classified. ACT analyses suggest that reducing number of tied ratings in response categories 5 and 6—both reflecting ambivalence—will boost the ESS significantly and involve a significant segment of the sample.

Table 4: Second ACT Omitting Ratings of 4-5 and 6-7

| | | Predicted Symptom | | |
|---|---|---|---|---|
| | | Pain | Fatigue | |
| Actual | Pain | 74 | 8 | 90.2% |
| Symptom | Fatigue | 5 | 72 | 93.5% |
| | | 93.7% | 90.0% | |

Table 5: Final ACT Using Ratings of 2 and 9

| | | Predicted Symptom | | |
|---|---|---|---|---|
| | | Pain | Fatigue | |
| Actual | Pain | 26 | 1 | 96.3% |
| Symptom | Fatigue | 0 | 23 | 100% |
| | | 100% | 95.8% | |

In summary UniODA comparison of raw pain and fatigue ratings made by this FM patient suggests that the latter dominate the former with moderate strength, and this finding may cross-generalize. Ratings that reflect ambivalence induce inaccurate classification by the model and the patient, and represent a possible means to enhance precision and thus accuracy.

## Analysis of Ipsative *z*-Scores

Ipsative standardization of raw data into *z*-score form is done using the mean and SD computed for data from an observation (not a sample of observations), and it is appropriate for analysis of serial data as a means of eliminating variability attributable to "base-rate" differences between observations that essentially introduce noise into the data.[3] Analyses performed on raw data are thus repeated here after pain and fatigue ratings have first been ipsatively standardized.

2155-0182/10/$3.00

Descriptive statistics for $z_{pain}$: mean=0; SD=1; median=0.265; skewness=0.20; kurtosis =-0.53; CV=0. For $z_{fatigue}$: mean=0; SD=1; median=-0.261; skewness=-0.03; kurtosis=-0.59; CV=0. Table 6 gives the $z_{pain}$ and $z_{fatigue}$ rating distributions for this patient.

Table 6: Distributions of the Patient's Ipsative $z_{pain}$ and $z_{fatigue}$ Ratings

| Response Category | $z_{pain}$ | $z_{fatigue}$ |
|---|---|---|
| -2.34 | | 5 |
| -1.70 | 26 | |
| -1.65 | | 25 |
| -1.05 | 48 | |
| -0.95 | | 50 |
| -0.39 | 74 | |
| -0.26 | | 85 |
| 0.27 | 65 | |
| 0.43 | | 60 |
| 0.92 | 49 | |
| 1.13 | | 49 |
| 1.58 | 27 | |
| 1.82 | | 23 |
| 2.23 | 7 | |
| 2.89 | 1 | |

Comparison of Tables 6 and 1 clearly reveals that "identical" rating scales, such as Likert-type scales, are definitely NOT identical if considered statistically from the perspective of the individuals who are using the scales to rate their own personal experience. This is in some extent due to ambiguity in the cognitive labels used to give meaning to the numerical options on the scale. It is also due to ambiguity in the target of the rating, the changing nature of the phenomenon over time, and the inherent differences in mean intensity and variability of rated phenomena. Little of this complexity is reflected in any systematic way with raw data, but this complexity is the theoretical motivation for the use of ipsatized data.[7,8]

The analysis begins as before: construct a new data set (*new.txt*) having 2*n* observations (here, 594), each forming a row in new.txt. Copy all 297 $z_{pain}$ ratings to new.txt (at this point new.txt has 297 rows), then beneath these copy all 297 $z_{fatigue}$ ratings (new.txt now has 594 rows). Add the number "0" (an arbitrary class category dummy-code) delimited by a space to the beginning of *each* of the *first* (top) set of 297 rows, and then likewise add the number "1" (arbitrary dummy-code) delimited by a space to the beginning of *each* of the *second* (bottom) set of 297 rows. Using new.txt as input, the *post hoc* hypothesis that $z_{pain}$ and $z_{fatigue}$ were rated differentially is tested by running the same UniODA code used for raw score analysis with the following changes:

```
VARS class Zrating;
ATTR Zrating;
```

The model had strong statistical support, with nearly perfect confidence for target $p<0.001$ (estimated $p<0.00001$). Analysis was completed in 42 CPU seconds by UniODA[4] run on a 3 GHz Intel Pentium D microcomputer.

The UniODA model was: if ipsative *z*-score $\leq$-0.33 (one-third of a SD lower than the mean rating used by the individual) then predict the attribute is *pain*; otherwise predict *fatigue*. As was found for raw data, using standardized data UniODA reveals that standardized fatigue ratings dominated pain ratings. The confusion table for this model is seen in Table 7: overall, ESS=22.9 (49% lower than obtained using raw data) and ESP=24.2 (46% lower than for raw data), reflecting relatively weak effects.[4] Model performance was stable in LOO analysis. Note that accuracy declined for classification of $z_{pain}$.

2155-0182/10/$3.00

Table 7: Confusion Table
for UniODA Model Discriminating
Ipsative Pain and Fatigue Ratings

| | | Predicted Symptom | | |
|---|---|---|---|---|
| | | $z_{pain}$ | $z_{fatigue}$ | |
| Actual | $z_{pain}$ | 148 | 149 | 49.8% |
| Symptom | $z_{fatigue}$ | 80 | 217 | 73.1% |
| | | 64.9% | 59.3% | |

All possible ACTs were examined next. The first ACT analysis dropped the middle response category of 0.27 (see Table 6), yielding the ACT provided in Table 8 (89% of the total sample). Resultant ESS (36.9) and ESP (37.0) were 61% greater than achieved in full-range analysis, and both were moderate effects.[4]

Table 8: ACT for UniODA Model
Discriminating Ipsative Pain and Fatigue
Ratings, Omitting Middle Rating

| | | Predicted Symptom | | |
|---|---|---|---|---|
| | | $z_{pain}$ | $z_{fatigue}$ | |
| Actual | $z_{pain}$ | 148 | 84 | 63.8% |
| Symptom | $z_{fatigue}$ | 80 | 217 | 73.1% |
| | | 64.9% | 72.1% | |

The second ACT also dropped response categories -0.26 and 0.43 producing the ACT in Table 9: ESS (32.3, 12% *lower* than first ACT) and ESP (38.4, 4% higher) are moderate effects, 41% greater than full-range analysis—half the gain of the first ACT, and only 65% of the total sample is classified—only three-quarters of the number classified by the first ACT. This shows that greatest accuracy gains can be realized by resolving ambiguity on *pain* ratings, not fatigue.

Table 9: ACT for UniODA Model
Discriminating Ipsative Pain and Fatigue
Ratings, Omitting Three Middle Ratings

| | | Predicted Symptom | | |
|---|---|---|---|---|
| | | $z_{pain}$ | $z_{fatigue}$ | |
| Actual | $z_{pain}$ | 148 | 84 | 63.8% |
| Symptom | $z_{fatigue}$ | 72 | 160 | 69.0% |
| | | 67.3% | 65.6% | |

The third and in the present case final ACT additionally dropped response categories of -0.39 and 0.92 producing the ACT in Table 10: ESS (48.2, 31% higher than any other result) and ESP (48.9, 27% higher) sit on the border of being strong effects, but only 44% of the total sample is classified. Again, eliminating non-extreme pain ratings returns solid improvement in overall model performance.

Table 10: ACT for UniODA Model
Discriminating Ipsative Pain and Fatigue
Ratings, Omitting Five Middle Ratings

| | | Predicted Symptom | | |
|---|---|---|---|---|
| | | $z_{pain}$ | $z_{fatigue}$ | |
| Actual | $z_{pain}$ | 74 | 35 | 67.9% |
| Symptom | $z_{fatigue}$ | 30 | 122 | 80.3% |
| | | 71.2% | 77.7% | |

Further ACTs are not performed because the proportion of the sample classified is already less than half, and analyses indicated that pain ratings near the patient mean offer the promise of improving accuracy vis-à-vis more precise measurement.

An important concluding observation is that although the model achieved using raw data had superior performance when compared with

2155-0182/10/$3.00

the model analyzing ipsative data—in terms of comparing the patient's ratings of two attributes, the raw data model has lower psychological meaning—if statistically considered from the perspective of the rater—than does the ipsative data model. The identical physical response scale may be provided to and used by a single person to rate two or more attributes, however that does not imply that the scale has the same psychological meaning when applied by the person to rate different attributes. Attributes should thus be ipsatively standardized before being compared against each other for a single person, and before being agglomerated in sample-based analyses.[8,9]

## References

[1]Collinge WC, Soltysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: Initial design and alpha test. *Optimal Data Analysis, 1*, 163-175.

[2]Collinge W, Yarnold PR, Soltysik, RC (2013). Fibromyalgia symptom reduction by online behavioral self-monitoring, longitudinal single subject analysis and automated delivery of individualized guidance. *North American Journal of Medical Sciences, 5*, 546-553.

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

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

[5]Yarnold PR, Soltysik RC (2010). Precision and convergence of Monte Carlo Estimation of 2-category UniODA 2-tailed *p*. *Optimal Data Analysis, 1*, 43-45.

[6]Yarnold PR (2013). Standards for reporting UniODA findings expanded to include ESP and all possible aggregated confusion tables. *Optimal Data Analysis*, 2, 106-119.

[7]Yarnold PR (1992). Statistical analysis for single-case designs. In: FB Bryant, L Heath, E Posavac, J Edwards, E Henderson, Y Suarez-Balcazar, S Tindale (Eds.), *Social Psychological Applications to Social Issues, Volume 2: Methodological Issues in Applied Social Research***.** New York, NY: Plenum, pp. 177-197.

[8]Yarnold PR, Feinglass J, Martin GJ, McCarthy WJ (1999). Comparing three pre-processing strategies for longitudinal data for individual patients: An example in functional outcomes research. *Evaluation and the Health Professions, 22*, 254-277.

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

## Author Notes

E-mail: Journal@OptimalDataAnalysis.com

Mail: Optimal Data Analysis, LLC
1220 Rosecrans St., #330
San Diego, CA 9210

*ODA* Blog: http://odajournal.wordpress.com