Obtaining Meaningful Latent Class Segments with Ratings Data by Adjusting for Level, Scale and Extreme Response Styles

Jay Magidson
Statistical Innovations
www.statisticalinnovations.com

Presented at M3 Conference, U. of Connecticut, May 2014

Background

- Obtaining meaningful Latent Class (LC) segments from ratings data is difficult because of response style confounds.
- Without accounting for such confounds, the resulting segments (classes) may differ primarily in their rating response styles rather than their preferences.

Overview

This paper introduces extended LC segmentation models which classify individuals while adjusting simultaneously for confounds due to level, scale (non-differentiation) and extreme response styles.

The models were estimated with the syntax module of Latent GOLD 5.0.

Example: Food taste test (Kellogg's)

- **Products**: 15 crackers
- **Consumers**: n=157 (category users)
  - evaluated all products over three days
  - 9-point liking scale (dislike extremely➔like extremely)
  - completely randomized block design balanced for the effects of day, serving position, and carry-over
- **Object**: To determine if consumers could be segmented according to their cracker ***preferences*** so The Kellogg Company can customize crackers for each segment.

Standard LC cluster data layout

crackers.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

8 : AvgRtg 6.6

	ID	R#117	R#138	R#231	R#342	R#376	R#410	R#495	R#548	R#603	R#682	R#755	R#812	R#821	R#951	R#967	AvgRtg
1	1101	6	7	6	6	6	8	9	9	7	8	6	9	9	8	8	7.47
2	1102	8	7	6	6	9	7	9	9	4	9	3	6	7	9	7	7.07
3	1103	8	3	5	6	7	6	3	9	7	8	5	8	2	7	2	5.73
4	1104	4	2	3	2	8	6	7	5	2	7	4	7	6	7	6	5.07
5	1105	2	2	8	2	7	4	9	8	5	5	3	9	7	7	7	5.67
6	1106	3	7	2	2	3	6	6	7	8	8	1	7	4	6	6	5.07
7	1107	1	1	1	2	5	9	1	8	5	9	1	9	8	9	5	4.93
8	1108	2	2	2	7	9	9	9	6	8	7	8	7	9	6	8	6.60
9	1109	8	8	7	3	8	8	9	8	7	9	7	9	8	9	9	7.80
10	1110	6	4	4	2	8	7	9	8	7	8	5	7	8	8	5	6.40

Data View Variable View

SPSS Processor is ready

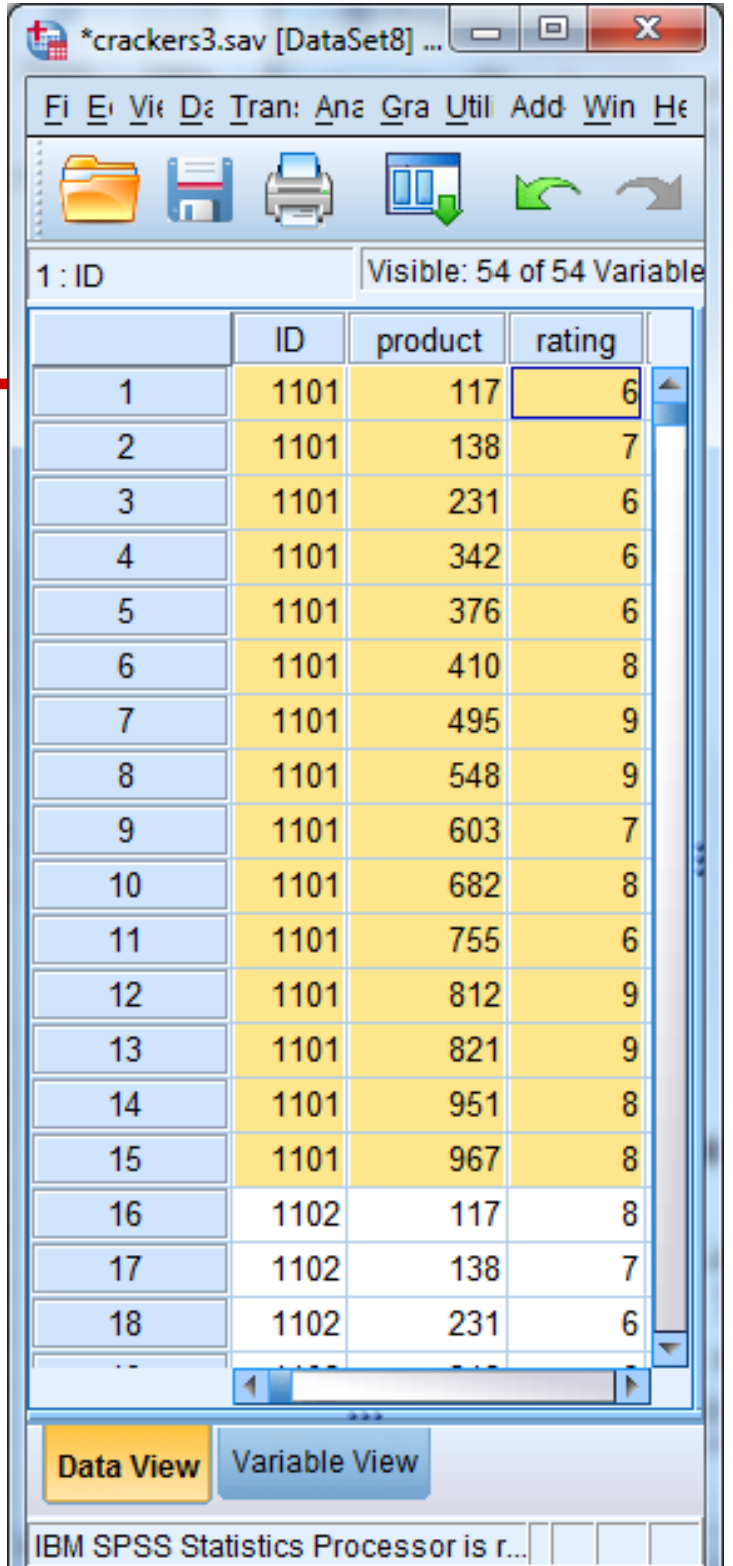
Ratings for each of the 15 products plus the average rating for each case

Long file format

This data format allows specification of various LC regression models with the ordinal variable RATING as the dependent variable and PRODUCT as the sole (nominal) predictor.

A) Simple LC Ordinal Logit Model:

$$log\left[\frac{P(Y_{it}=m\,|\,x)}{P(Y_{it}=m-1\,|\,x)}\right]=\alpha_m+\beta_x+\mu_t+\gamma_{xt}$$



	ID	product	rating
1	1101	117	6
2	1101	138	7
3	1101	231	6
4	1101	342	6
5	1101	376	6
6	1101	410	8
7	1101	495	9
8	1101	548	9
9	1101	603	7
10	1101	682	8
11	1101	755	6
12	1101	812	9
13	1101	821	9
14	1101	951	8
15	1101	967	8
16	1102	117	8
17	1102	138	7
18	1102	231	6

Simple LC ordinal logit model

$$log\left[\frac{P(Y_{it}=m\,|\,x)}{P(Y_{it}=m-1\,|\,x)}\right]=\alpha_m+\beta_x+\mu_t+\gamma_{xt}$$

Model Type A:
Assumes no confounds

where:

product $t = 1,2,\ldots,15$; x=1,2,…,K classes

α_m are the intercepts measuring the overall size (usage) of each ratings level m= 1,2,…,9

γ_{xt} is the effect for product t for cases in latent class x;

effect coding is used for parameter identification: e.g.,

$$\sum_{t=1}^{T}\mu_t=\sum_{x=1}^{K}\beta_x=\sum_{t=1}^{T}\gamma_{xt}=\sum_{x=1}^{K}\gamma_{xt}=0$$

(so the intercepts capture overall response levels)

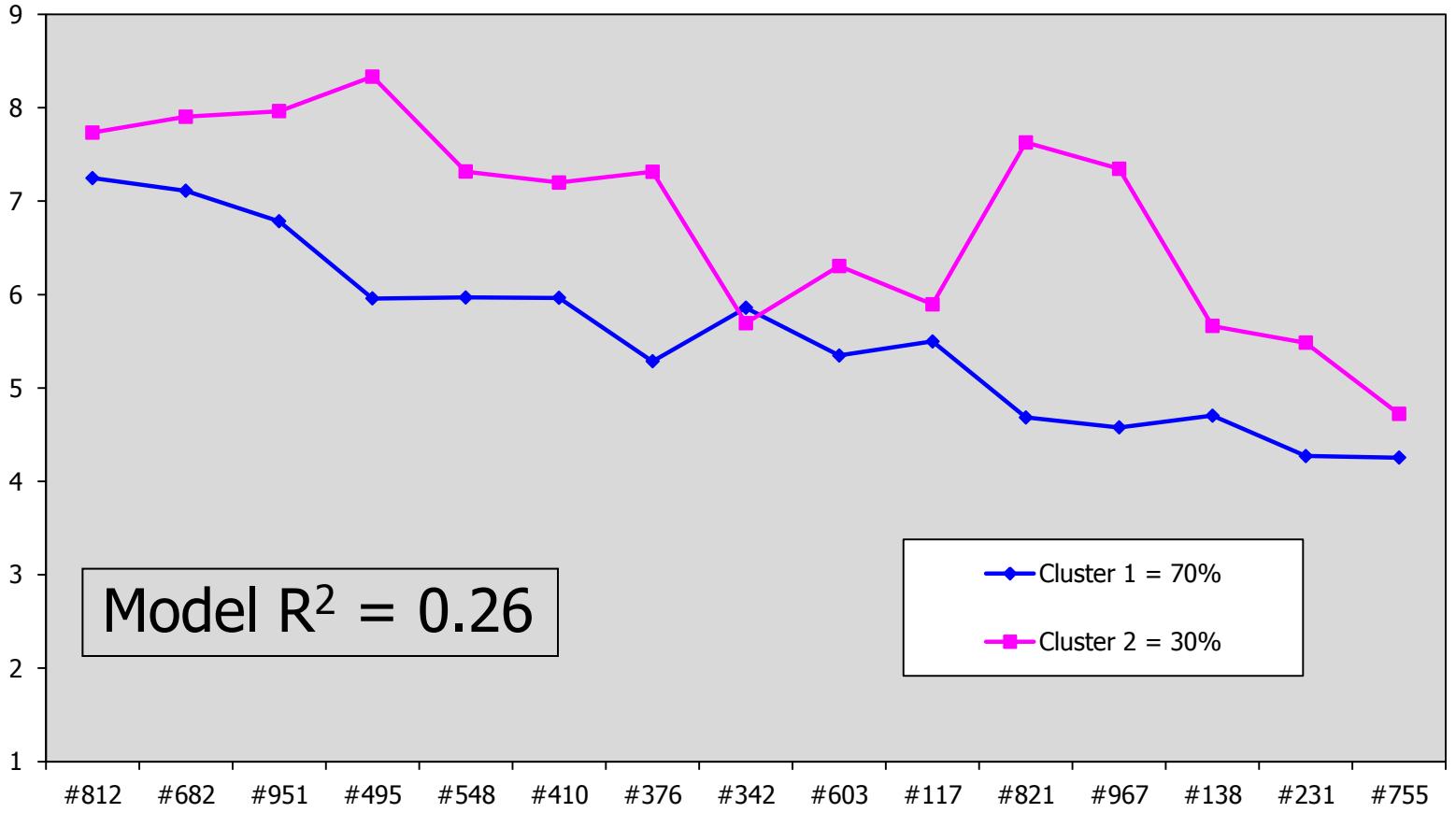
Results: BIC suggests K=2 segments

Model	Description	LL	BIC(LL)	Npar	R^2
A0	Null Model	-4955.8	9952.0	8	0
A1	Ordinal Regression	-4762.4	9636.1	22	0.15
A2	2-class Simple Ordinal Regression	-4682.4	9556.9	38	0.26
A3	3-class Simple Ordinal Regression	-4645.8	9564.6	54	0.32

Based on the BIC (Bayesian Information Criterion), a 2-class solution fit the data better than either 1-class or 3-class solutions.

The resulting segments are not useful to the food manufacturer (Kellogg's)

Segment 2 tends to rate all crackers higher than Segment 1.



Model B: Random intercept LC regression to account for response level (L) effects

$$logit(Y_{im.t}) = \alpha_{im} + \beta_x + \mu_t + \gamma_{xt}$$
$$\alpha_{im} = \alpha_m + \lambda F_i$$

Thus,

$$E(\alpha_{im}) = \alpha_m$$
$$V(\alpha_{im}) = \lambda^2$$

where: F_i is a continuous factor (CFactor)

$$F_i \sim N(0,1)$$ or equivalently,

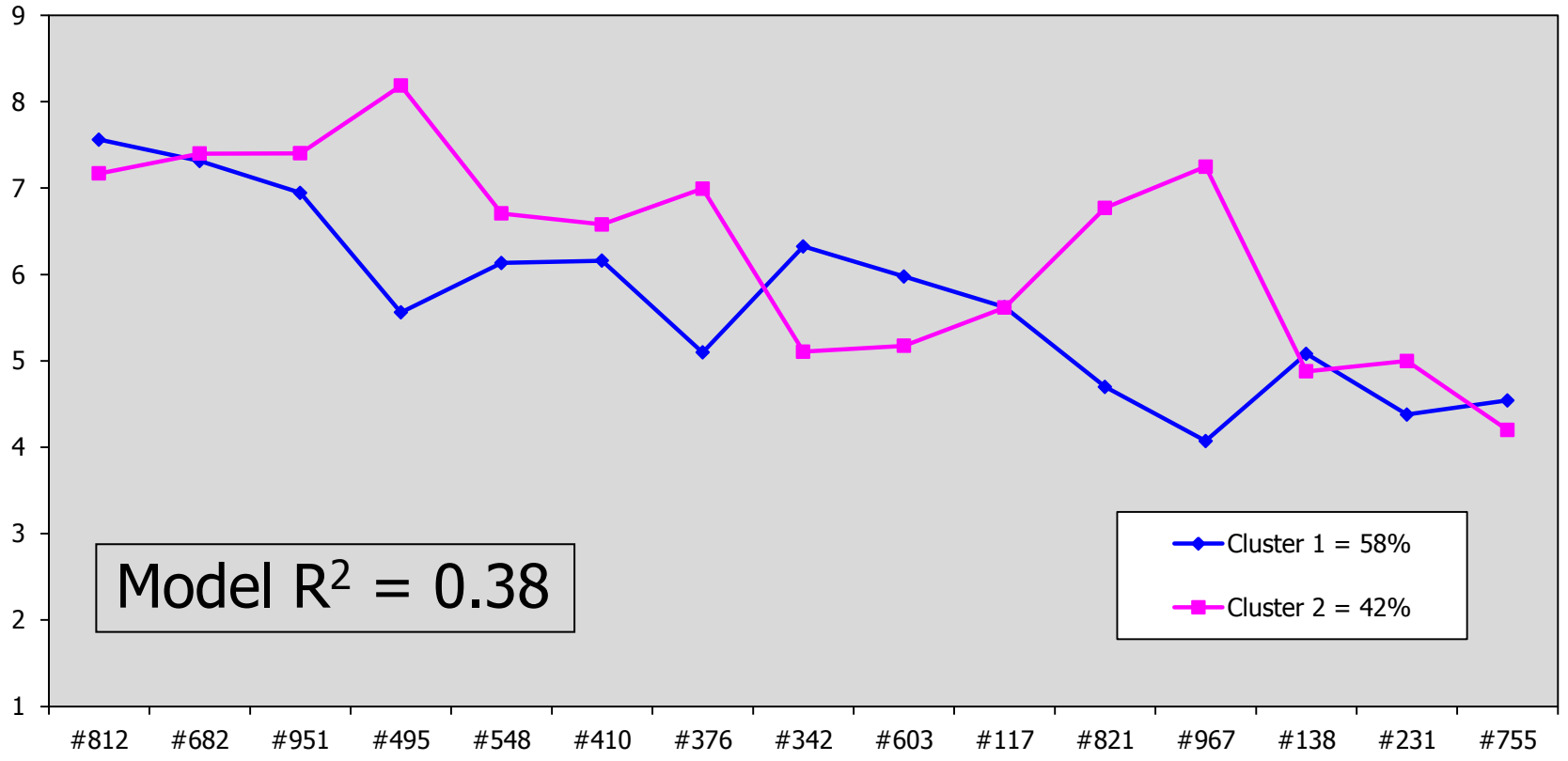
$$\alpha_{im} \sim N(\alpha_m, \lambda^2) \qquad \text{m} = 1,2,\ldots,9$$

BIC again suggests 2 segments when intercept is random

Model	Description	LL	BIC(LL)	Npar	R^2
	Null Model	-4955.8	9952.0	8	0
A1	Ordinal Regression	-4762.4	9636.1	22	0.15
A2	2-class Simple Ordinal Regression	-4682.4	9556.9	38	0.26
A3	3-class Simple Ordinal Regression	-4645.8	9564.6	54	0.32
B2	2-classes + Rand. Int.	-4641.4	9480.1	39	0.38
B3	3-classes + Rand. Int.	-4617.0	9512.2	55	0.41

This is an improvement over the simple 2-class LC regression model

Model B: Resulting segments are more meaningful



Model B: Use of random intercept is more appropriate than mean-centering

- Correlation of random intercept with average liking is 0.997
- Inclusion of random intercept is conceptually similar to mean-centering each respondents' liking ratings
  - LC Cluster model of the mean-centered data produces similar results
- Advantages of LC Regression over mean-centering
  - maintains ordinal metric
  - can be used with partial profile (incomplete block) designs

Other response style confounds in these data?

Extreme response style

Moore (2003) and Morren et al (2011) suggest use of one or more discrete factors (DFactors) of the kind introduced by Magidson and Vermunt (2001) to account for extreme (E-type) response styles:

$$log\left[\frac{P(Y_{it}=m\,|\,x,L_i,E_j)}{P(Y_{it}=m-1\,|\,x,L_i,E_j)}\right]=\alpha_m+\omega_m E_j+\lambda L_i+\beta_x+\mu_t+\gamma_{xt}$$

Scale (S-type) Heterogeneity – referred to as non-differentiation in survey research, we model this using the Vermunt (2013) specification:

$$log\left[\frac{P(Y_{it}=m\,|\,x,L_i,S_i,E_j)}{P(Y_{it}=m-1\,|\,x,L_i,S_i,E_j)}\right]=\exp(\lambda S_i)(\alpha_m+\omega_m E_j+\lambda L_i+\beta_x+\mu_t+\gamma_{xt})$$

Are there other response style confounds?

Model fit is improved when E effect terms are added (as either discrete or continuous)

$$log\left[\frac{P(Y_{it}=m\,|\,x,L_i,E_j)}{P(Y_{it}=m-1\,|\,x,L_i,E_j)}\right]=\alpha_m+\omega_m E_j+\lambda L_i+\beta_x+\mu_t+\gamma_{xt}$$

E as DFactor: j=1,2,…,J

$$log\left[\frac{P(Y_{it}=m\,|\,x,L_i,E_i)}{P(Y_{it}=m-1\,|\,x,L_i,E_i)}\right]=\alpha_m+\omega_m E_i+\lambda L_i+\beta_x+\mu_t+\gamma_{xt}$$

E modeled as CFactor

Are there other response style confounds?

Fit is further improved when the scale factor (S) term is added (shown as continuous below):

$$log\left[\frac{P(Y_{it}=m\,|\,x,L_i,S_i,E_j)}{P(Y_{it}=m-1\,|\,x,L_i,S_i,E_j)}\right]=\exp(\lambda S_i)(\alpha_m+\omega_m E_j+\lambda L_i+\beta_x+\mu_t+\gamma_{xt})$$

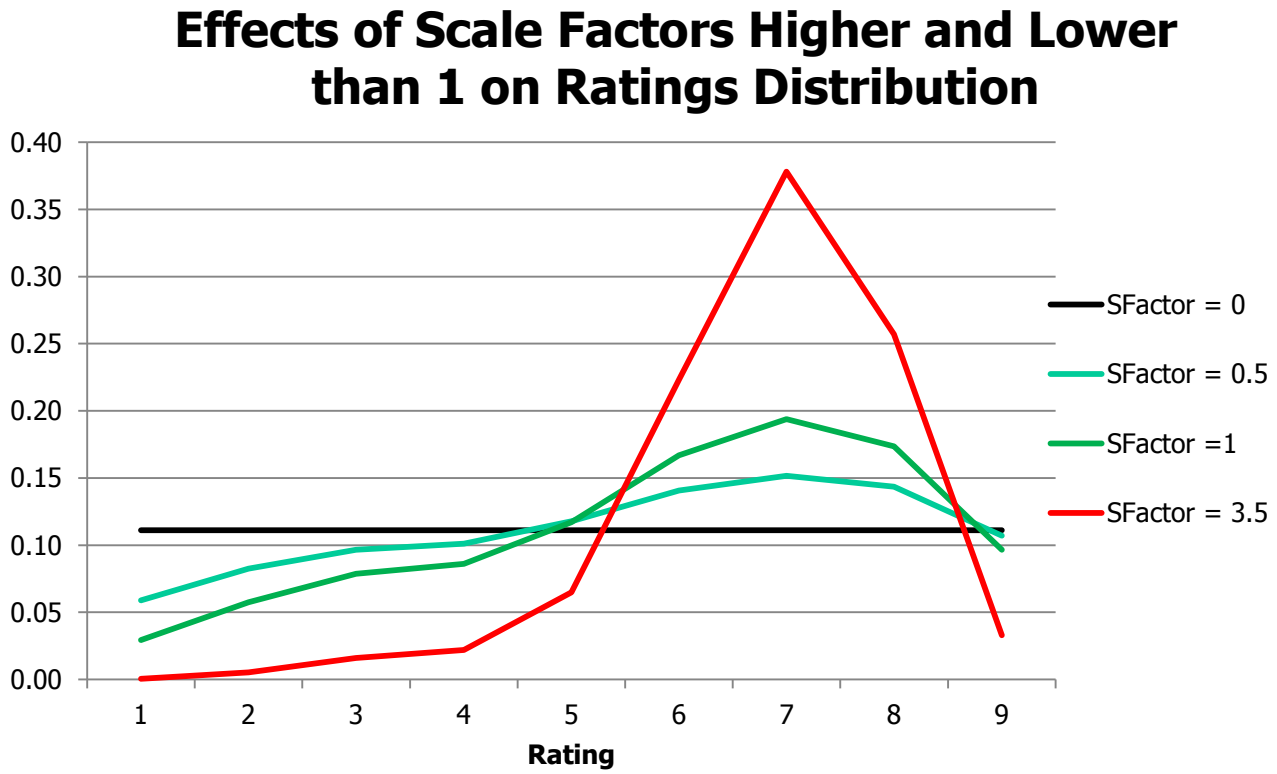
$$log\left[\frac{P(Y_{it}=m\,|\,x,L_i,S_i,E_i)}{P(Y_{it}=m-1\,|\,x,L_i,S_i,E_i)}\right]=\exp(\lambda S_i)(\alpha_m+\omega_m E_i+\lambda L_i+\beta_x+\mu_t+\gamma_{xt})$$

Model fit improves further when accounting for L, E and S confounds

Model	Description	LL	BIC(LL)	Npar	R^2
	Null Model	-4955.8	9952.0	8	0
A1	Ordinal Regression	-4762.4	9636.1	22	0.15
A2	2-class Simple Ordinal Regression	-4682.4	9556.9	38	0.26
A3	3-class Simple Ordinal Regression	-4645.8	9564.6	54	0.32
B2	2-classes + Rand. Int.	-4641.4	9480.1	39	0.38
B3	3-classes + Rand. Int.	-4617.0	9512.2	55	0.41
E2c	2-class(L=c, E=c, S=c) with no E-restrictions	-4567.1	9376.8	48	0.41
E2DFac3	2-class(L=c, E=DFac3, S=c) with no E-restrictions	-4566.4	9385.5	50	0.41

Model E2c uses a continuous latent factor to model the E effect

Scale factors stretch or shrink the distribution: SFactor = $\exp(\lambda S_i)$

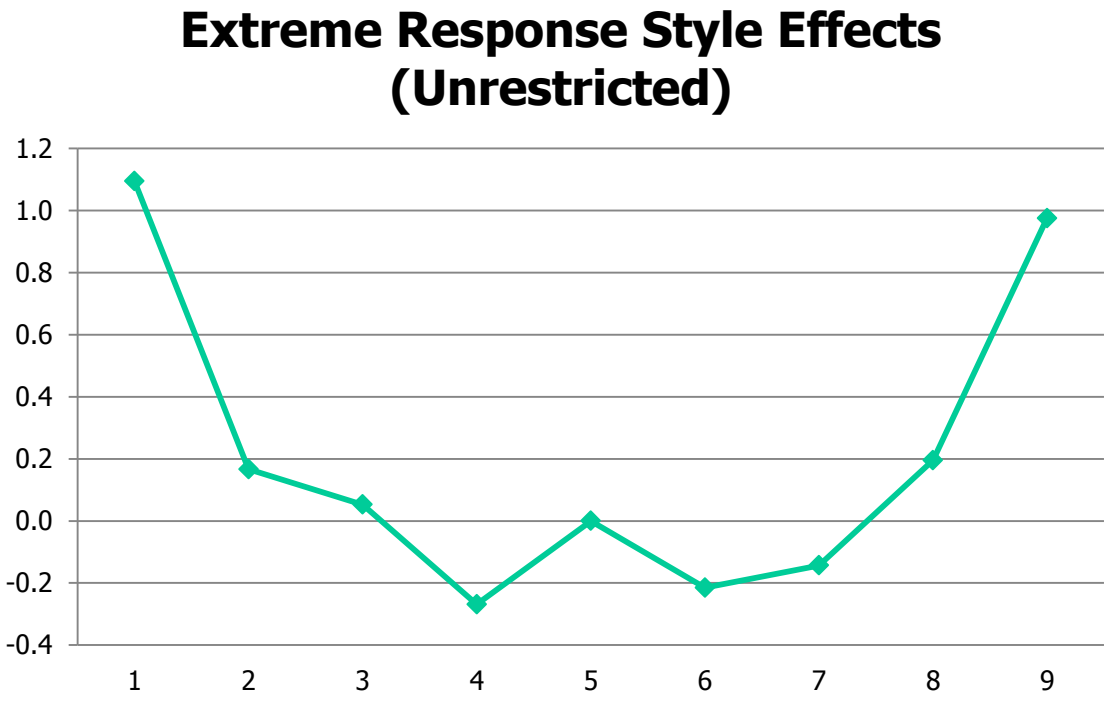


Scale factors greater than 1 stretch the distribution (ratings 6-8 become more likely) while those less than 1 shrink towards the uniform distribution (all ratings 1-9 become equally likely).

The E effects capture the extreme response style

A plot of the ω estimates for model E2c indicates preference for the extreme responses over the middle level (set to 0 for identification) for persons with a high CFactor score Ei.

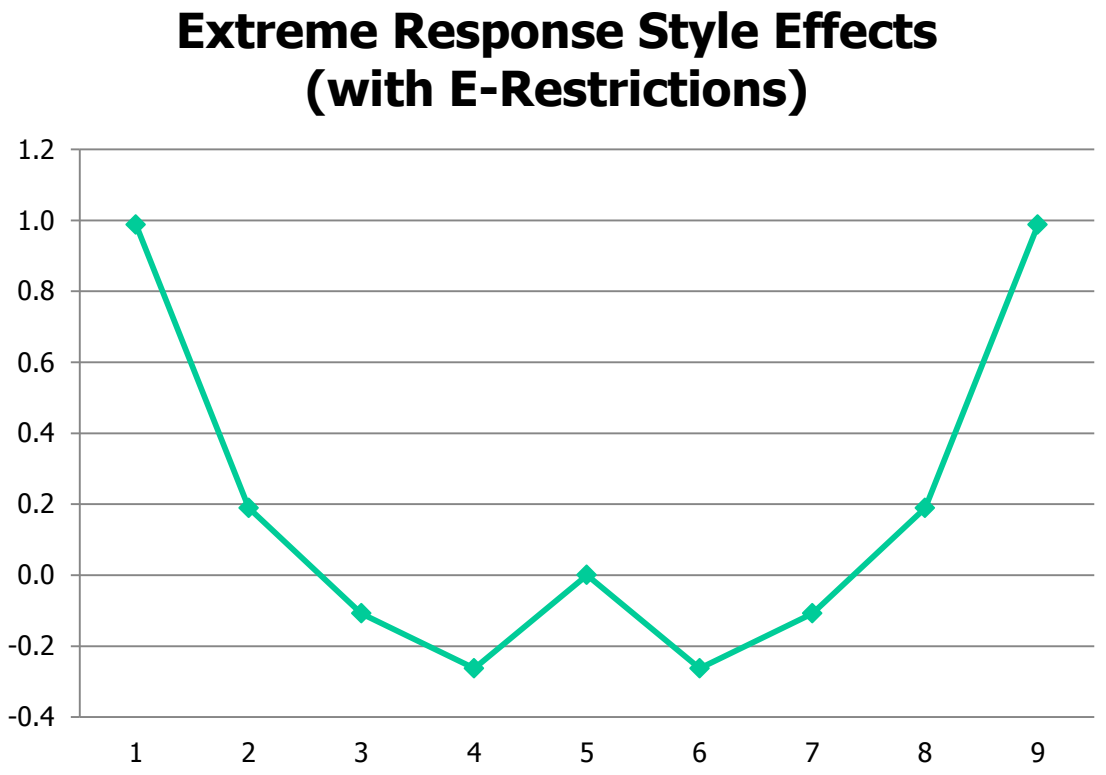
Also, levels 4 and 6 are somewhat less preferred than the midlevel, indicating a small midlevel response style.



Symmetry restrictions for extreme response style

Since there is no guarantee that the E-effects will capture the extreme response style (ERS), symmetry restrictions can be applied which will tend to identify ERS if these restrictions fit:

$$\omega_9 = \omega_1$$
$$\omega_8 = \omega_2$$
$$\omega_7 = \omega_3$$
$$\omega_6 = \omega_4$$



Symmetry restrictions are supported by the data

Model	Description	LL	BIC(LL)	Npar	R^2
	Null Model	-4955.8	9952.0	8	0
A1	Ordinal Regression	-4762.4	9636.1	22	0.15
A2	2-class Simple Ordinal Regression	-4682.4	9556.9	38	0.26
A3	3-class Simple Ordinal Regression	-4645.8	9564.6	54	0.32
B2	2-classes + Rand. Int.	-4641.4	9480.1	39	0.38
B3	3-classes + Rand. Int.	-4617.0	9512.2	55	0.41
E2c	2-class(L=c, E=c, S=c) with no E-restrictions	-4567.1	9376.8	48	0.41
E2DFac3	2-class(L=c, E=DFac3, S=c) with no E-restrictions	-4566.4	9385.5	50	0.41
E2c_rest	2-class(L=c, E=c, S=c) with E-restrictions	-4568.6	9359.7	44	0.41

Symmetry restrictions were also applied in a reanalysis of data used in Morren, et. al. (2011) and found to fit those data better than the Morren et. al approach for modeling extreme response styles.

Classification is likely to improve substantially when accounting for response styles

Model	Class size 1	2
Simple 2-class ordinal regression	70%	30%
Adjusting for Level (L) effect	58%	42%
Adjusting for L, E and S	52%	48%

Class 1* consists of 70% of the raters (with standard error .08) in the first model vs. 57.5% (with standard error .07) in the final model.

Prediction also improves: R^2 increases from .26 to .41

* Class 1 has the lower overall mean rating

Summary

We compared various LC models designed to identify meaningful segments that differed in their *preferences* (not just ratings) for crackers. After adjusting for level, extreme rating and scale heterogeneity we obtained a model that fit substantially better and was more meaningful and useful to the food manufacturer.

In this application we also introduced a new approach for modeling the extreme response styles by applying symmetry restrictions.

Notes

The symmetry restrictions were also applied in a reanalysis of data used in Morren, et. al. (2011) and found to fit substantially better than the Morren et. al approach.

For further analysis of these data using product attributes, see Popper et. al (2004).

References

Magidson, J., J.K. Vermunt. (2001). “Latent class factor and cluster models, bi-plots and related graphical displays”. *Sociological Methodology*, 31, 223-264

Magidson, J., J.K. Vermunt. (2006). “Use of latent class regression models with a random intercept to remove overall response level effects in ratings data”. In A. Rizzi and M Vichi (eds.), *Proceedings in Computational Statistics* , 351-360, Heidelberg: Springer.

Magidson, J., J.K. Vermunt. (2013). Latent GOLD syntax manual and Latent GOLD 5.0 upgrade manual. Statistical Innovations: Belmont Massachusetts.

Moors, G. (2003). Diagnosing Response Style Behavior by Means of a Latent-Class Factor Approach. Socio-Demographic Correlates of Gender Role Attitudes and Perceptions of Ethnic Discrimination Reexamined. Quality & Quantity, 37(3), 277-302.

Moors, G. (2008) “Exploring the effect of a middle response category on response style in attitude measurement”, Quality and Quantity December; 42(6): 779–794.

Morren, M., Gelissen, J.P.T.M., and Vermunt, J.K. (2011). Dealing with extreme response style in cross-cultural research: A restricted latent class factor analysis approach. Sociological Methodology , 41, 13-47.

Popper, R., J Kroll, and J. Magidson (2004) Applications of latent class models to food product development: a case study. *Sawtooth Software Proceedings*, 2004.

Vermunt, J.K. (2013). Categorical response data. In: M.A. Scott, J.S. Simonoff, and B.D. Marx (eds.), The SAGE Handbook of Multilevel Modeling, 287-298. Thousand Oaks, CA: Sage.

Acknowledgment

The authors wish to thank The Kellogg Company for providing the data for this case study.

Latent GOLD 5.0 syntax for models

Null Model

- variables
- caseid ID;
- dependent rating;
- independent product nominal;
- equations
- rating <- 1 + (0)product ;
- // mean rating constant for all products and all cases

Latent GOLD Syntax

Model A1: Ordinal Regression

- variables
- caseid ID;
- dependent rating;
- independent product nominal;
- equations
- rating <- 1 + product ;
- // mean rating constant for all cases

Latent GOLD Syntax

Model A2: 2-class Simple Ordinal Regression

- variables
- caseid ID;
- dependent rating;
- independent product nominal;
- latent
- Class nominal 2;
- equations
- Class <- 1;
- rating <- 1 + Class + product | Class;
- // or rating <- 1 + Class + product + product * class;

Latent GOLD Syntax

Model B2: 2-classes + Random Intercept

- variables
- caseid ID;
- dependent rating;
- independent product nominal;
- latent
- Cfactor1 continuous,
- Class nominal 2;
- equations
- (1)Cfactor1;
- Class <- 1;
- rating <- 1 + Class + Cfactor1 + product | Class;

Latent GOLD Syntax

Model 2-class(L=c, E=c, S=c) with E-restrictions

- variables
- caseid ID;
- dependent rating coding = 5;
- independent product nominal;
- latent
- cfac continuous,
- cfac2 continuous , scfac continuous,
- Class nominal 2;
- equations
- (1)cfac ; (1) cfac2; scfac;
- Class <- 1;
- rating <- 1 + cfac+ (a~nom)Cfac2 + Class + product + product Class;
- rating <<- (1)scfac;
- a[1,8]=a[1,1]; a[1,7]=a[1,2]; a[1,6]=a[1,3]; a[1,5]=a[1,4];