

**Diagnostics for Frequency Models based on the
Two-Point Mixture Index of Model Fit**

C. Mitchell Dayton
Department of Measurement & Statistics
University of Maryland

Paper prepared for the Psychometric Society Annual Meeting
UNC Chapel Hill, June 2002

Two-Point Mixture Index of Fit

Let P be the “true” distribution for the cell proportions in a frequency table. Rudas, Clogg and Lindsay (1994) {RCL hereafter} propose a two-point mixture model:

$$P = (1 - \pi) \cdot \Phi + \pi \cdot \Psi \quad [1]$$

where

$\Phi =$ probability distribution implied by probabilistic model, H

$\Psi =$ an arbitrary, unspecified probability distribution

$0 \leq \pi \leq 1 =$ proportion of the population not consistent with H

Comparison with Pearson Chi-Square Statistic

For the two-point mixture model, the “expected” proportions, associated with H are always equal to or less than corresponding observed proportions [$\hat{P}_{ij} \leq P_{ij}$].

Note that this representation is different than the usual “fit” and “lack-of-fit” components associated with Pearson chi-square goodness-of-fit procedures. For example, for a two-way frequency table, let E_{ij} represent theoretical expected frequencies based on, say, an independence model. In general, E_{ij} may be less than, equal to or greater than O_{ij} .

Comparison with Goodman Intrinsically Unscalable Latent Class Scaling Model

Goodman (1975) proposed a latent class model for scaling (e.g., linear Guttman scale) in which each scale type was identified with a separate latent class and, in addition, one or more classes representing unscalable respondents were included. Dayton & Macready (1980) extended the Goodman model by incorporating various types of measurement errors for the scalable types. In a typical application, for example, it may be found that a hypothetical model is estimated to have 85% of respondents associated with various scale types but 15% of respondents do not fall into these types and are classified as intrinsically unscalable.

The RCL two-point mixture model may be viewed as a generalization of the Goodman model to a wider range of frequency data tables.

Definition of the Fit Index

π in Equation [1] is not unique and the equation is true, *de facto*, for any model for any frequency table.

The index of fit, π^* , is defined as the **smallest** value of π for which the representation in Equation [1] holds:

$$\pi^* = \inf\{\pi \mid P = (1-\pi) \cdot \Phi + \pi \cdot \Psi, \Phi \in H\} \quad [2]$$

π^* can be interpreted as the minimum proportion of cases that must be omitted from the frequency table in order to provide perfect fit to the remaining data.

Properties of the Fit Index

Assume that $\hat{\pi}^*$ is the maximum likelihood estimator of π^* .

Then,

$\hat{\pi}^*$ is unique

$\hat{\pi}^*$ is defined on the 0,1 interval

for nested models, $\hat{\pi}^*$ has the property of decreasing (actually, never increasing) in magnitude for increasingly more complex models

$\hat{\pi}^*$ is invariant if frequencies in a contingency table are increased/decreased by an arbitrary multiplicative factor

Held Back in School Crosstabulated with Sex 5% Sample of NELS Data

Frequencies			
	Female	Male	
No	498 (.85)	468 (.81)	
Yes	88 (.15)	109 (.19)	
Sum	586	577	N = 1163

$$\chi^2 = 3.10 \qquad p = 0.078$$

	Female	Male	
No	498.0	468.0	
Yes	88.0	82.7	
Sum	586.0	550.7	N* = 1136.7

$$\hat{\pi}^* = 0.023$$

Guidelines (?)

There is no general guideline for interpreting the two-point mixture index but, intuitively, values of 10% to 5% or less seem small. RCL remark that 10% is “reasonable” for a specific 4x4 cross-classification table but there is no absolute standard for the index that represents acceptable fit in all settings. In particular, 10% for the first example in the RCL paper represents only about 59 respondents whereas it represents about 2526 respondents for their second example.

Estimation of π^*

(1) Stepwise Search - RCL describe a stepwise computational approach for two-way frequency tables that can be generalized to frequency tables in general. In brief, they fix π^* at a suitably small value such as .01, solve for the components of the two-point mixture using an EM algorithm, increment π^* by, say, .01, and repeat estimation and incrementing until a saturated model is attained (i.e., until the G^2 goodness of fit statistic becomes 0). This approach is outlined by RCL, implemented in their FORTRAN program, Mixit, and described in detail by Xi (1994).

(2) Nonlinear Programming (NLP) approach - Xi (1994) and Xi & Lindsay (1996) conceptualize the estimation problem for $\hat{\pi}^*$ in terms of optimization using nonlinear programming (NLP) techniques. In particular, by decomposing frequencies in the model in equation [1] into components due to $\Phi \in H$, representing fit, and due to Ψ , representing lack of fit, Xi (1994) shows that NLP provides identical estimates to the corresponding EM algorithm.

NLP Computing Steps for Frequency Data

(a) Provide start values for the parameters $\theta = \{\theta_m\}$ and include the sum of

expected frequencies, $\sum_{j=1}^J n_j^* \leq N$ as a parameter for the NLP algorithm.

(b) Define expected frequencies for the cells of the frequency table

using the model, H, based on θ .

(c) Impose the restrictions $n_j^* \leq n_j$ for all j; in addition, impose

relevant restrictions on the parameters (e.g., non-negativity); in

Microsoft Excel Solver, these are termed *Constraints*.

(d) Define the objective function to be maximized as the sum of the

expected frequencies, $\sum_{j=1}^J n_j^*$; in Microsoft Excel Solver, this is

called the *Target Cell*; at convergence of the NLP algorithm,

$$\hat{\pi}^* = 1 - \sum_{j=1}^J n_j^* / N$$

(e) NOTE: These steps can be implemented using, for example, Excel

Solver or Gauss sqpsolve routine. See Dayton (1999, 2002) for

more details.

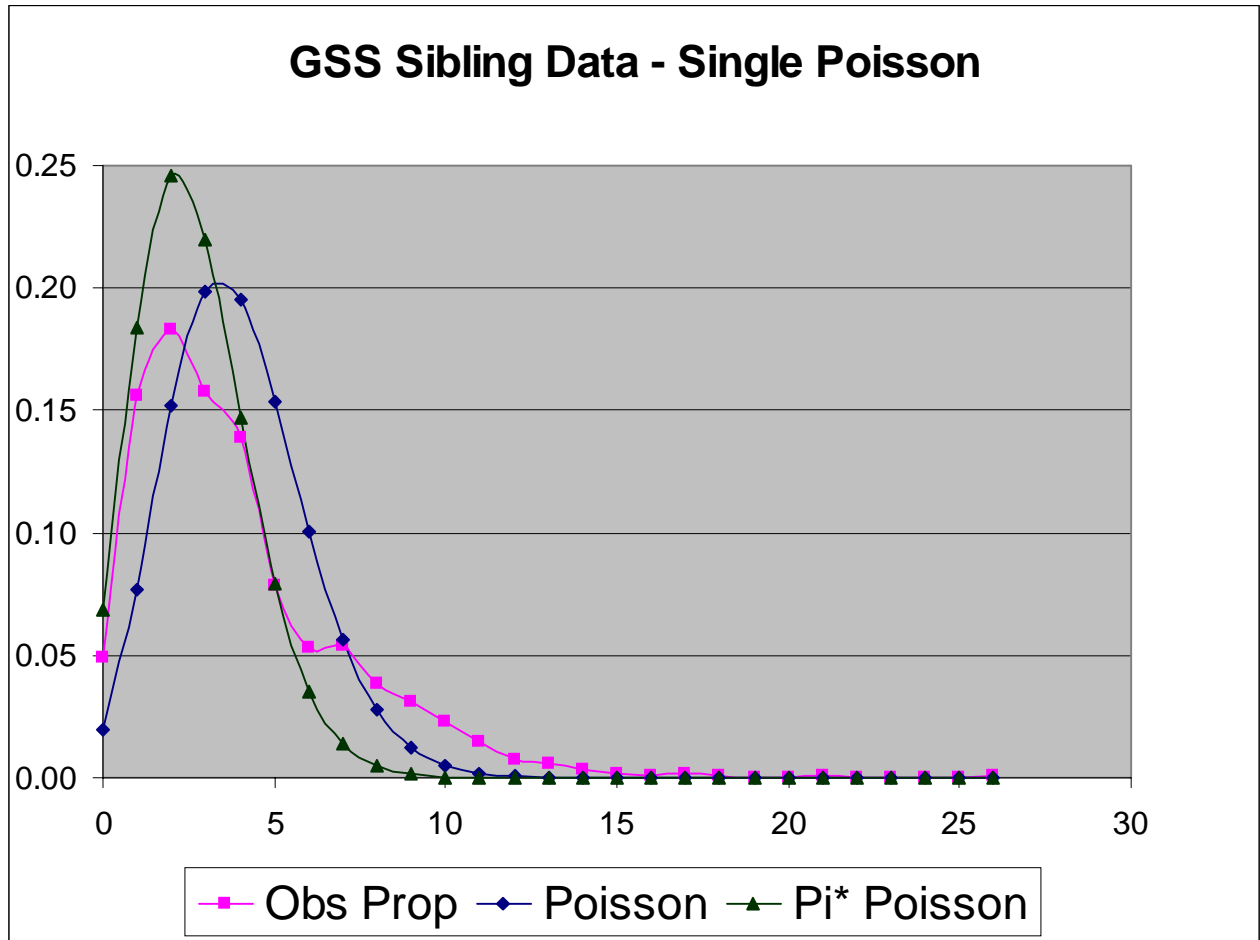
Estimating a Lower Bound for the Two-Point Mixture Index

The estimate, $\hat{\pi}^*$, is subject to random fluctuation due to the peculiarities of sample data. In general, $\hat{\pi}^*$ may overestimate lack of fit. RCL derived a lower confidence bound, $\hat{\pi}_L$, based on a G^2 fit statistic equal to 2.70 (i.e., the 90th percentage point of the chi-square distribution with one degree of freedom). Their program, Mixit, can be used to find the lower limit by the same iterative procedure used to compute $\hat{\pi}^*$. The confidence interval is one-sided since all values of $\hat{\pi}$ greater than $\hat{\pi}^*$ yield models of the form of equation [1] that fit the observed frequencies perfectly (i.e., $G^2 = 0$ if $\hat{\pi} > \hat{\pi}^*$). For more general data situations than those that can be fit by Mixit, the standard error of $\hat{\pi}^*$ can be estimated using re-sampling techniques (e.g., the jackknife; see Dayton, 1999, Dayton 2002). Clogg, Rudas & Xi (1995) suggest that the difference, $\hat{\pi}^* - \hat{\pi}_L$, provides a measure of the effect of sample size on the estimator, $\hat{\pi}^*$.

GSS Number of Siblings Data
Single Poisson Process

# Sibs	Observed		Single Poisson		Two-Point Mixture	
	Freq	Prop	E(Prop)	E(Freq)	E(Prop)	E(Freq)
0	74	0.049	0.020	29.56	0.069	73.89
1	235	0.156	0.077	116.19	0.184	198.02
2	276	0.183	0.152	228.30	0.246	265.35
3	237	0.157	0.199	299.08	0.220	237.04
4	209	0.139	0.195	293.84	0.147	158.82
5	118	0.078	0.153	230.96	0.079	85.13
6	80	0.053	0.101	151.28	0.035	38.02
7	81	0.054	0.056	84.93	0.014	14.56
8	58	0.039	0.028	41.72	0.005	4.88
9	47	0.031	0.012	18.22	0.001	1.45
10	34	0.023	0.005	7.16	0.000	0.39
11+	56	0.037	0.002	3.74	0.000	0.12
	1505	1.000	1.000	1505.00	1.000	1077.66

$G^2 = 586.962$	$\pi^* = 0.284$	$\pi^*_L = 0.221$
$\lambda = 3.93$	$\lambda = 2.68$	



NOTE: Observed proportions and single Poisson process are based on $N = 1505$ whereas RCL π^* parameter estimate is based on $N^* = 1077.66$.

GSS Number of Siblings Data
Mixture of Two Poissons

# Sibs	Observed		Mixture 2 Poissons		Two-Point Mixture	
	Freq	Prop	E(Prop)	E(Freq)	E(Prop)	E(Freq)
0	74	0.049	0.054	80.70	0.054	74.00
1	235	0.156	0.142	213.84	0.142	195.76
2	276	0.183	0.190	285.35	0.189	260.94
3	237	0.157	0.172	259.13	0.172	236.99
4	209	0.139	0.124	186.69	0.124	171.21
5	118	0.078	0.082	123.04	0.082	113.58
6	80	0.053	0.057	85.98	0.058	80.00
7	81	0.054	0.045	67.96	0.046	63.42
8	58	0.039	0.038	57.11	0.038	53.11
9	47	0.031	0.031	46.85	0.031	43.25
10	34	0.023	0.024	35.88	0.024	32.84
11+	56	0.037	0.042	62.46	0.041	56.01
	1505	1.000	1.000	1505.00	1.000	1381.10

$G^2 = 11.221$

$\lambda = 2.64, 7.81$

$\theta = .75, .25$

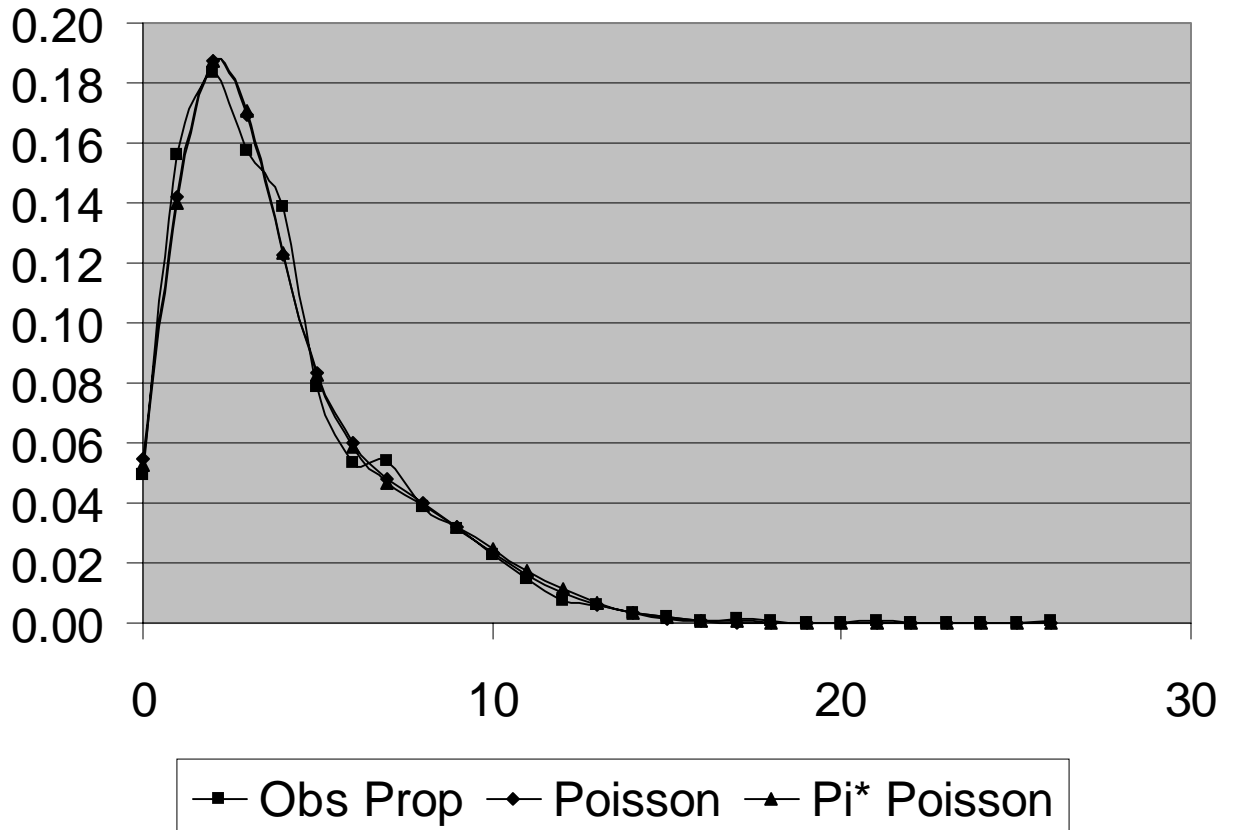
$\pi^* = 0.081$

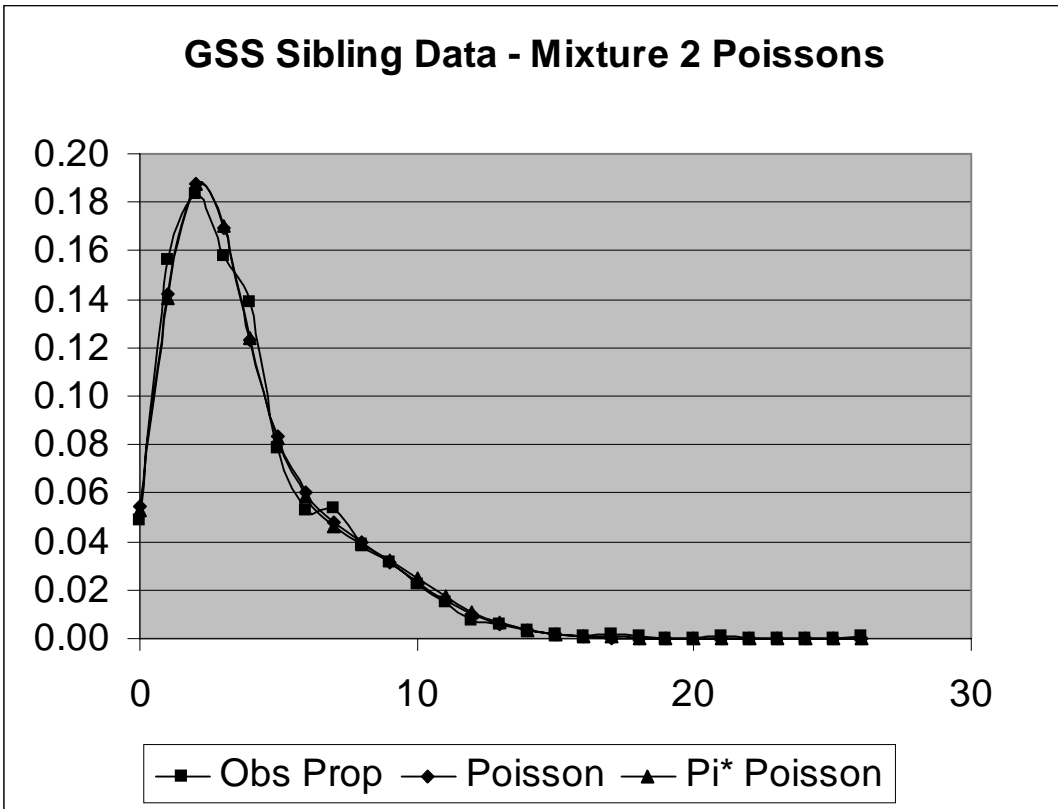
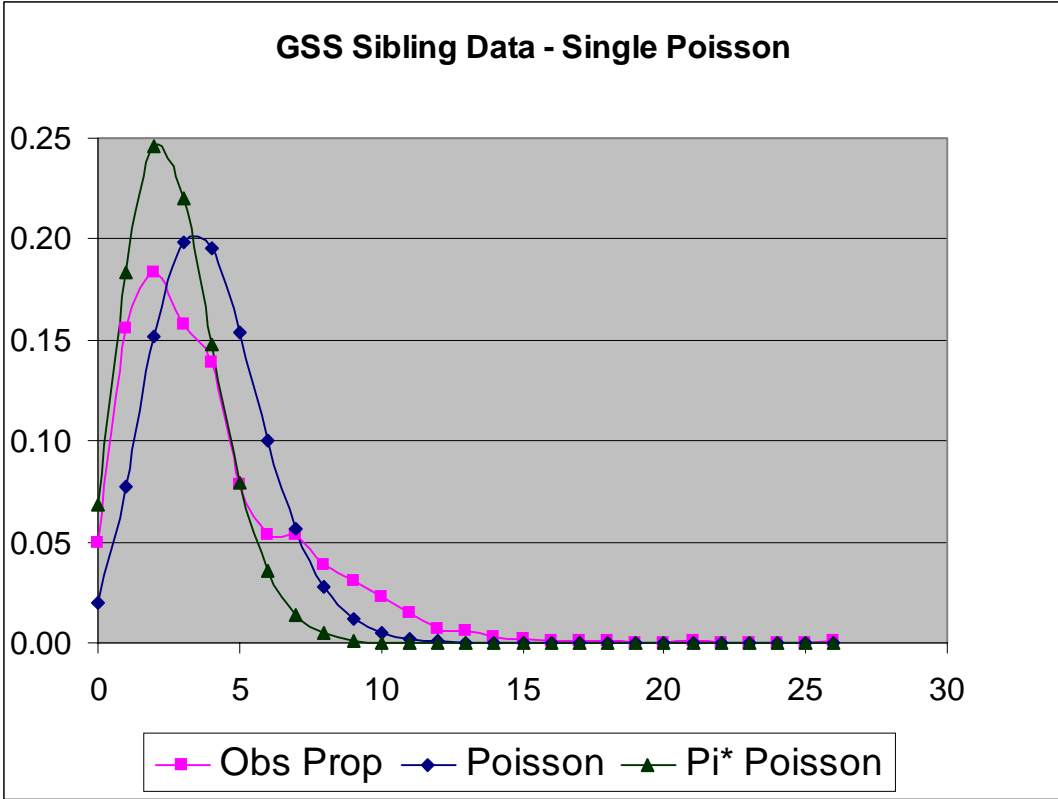
$\lambda = 2.63, 7.74$

$\theta = .75, .25$

$\pi^*_L = 0.019$

GSS Sibling Data - Mixture 2 Poissons





MSPAP - π^* for Rasch (1PL) Model

First 5 items from 17 item set

A	B	C	D	E	Freq	Prob [*]	E(F [*])	Standard Solution		π^* Solution	
								θ	δ	θ^*	δ^*
0	0	0	0	0	1614	0.141	1614.000	-1.88	1.42	-2.17	1.55
1	0	0	0	0	594	0.026	300.740	-1.15	0.81	-1.06	0.71
0	1	0	0	0	375	0.011	129.496	-0.43	0.30	-0.33	0.34
1	1	0	0	0	262	0.010	113.109	0.59	2.40	0.72	3.01
0	0	1	0	0	89	0.008	89.000	2.49	0.0	2.63	0.0
1	0	1	0	0	109	0.007	77.738	-0.35		-0.38	
0	1	1	0	0	50	0.003	33.473				
1	1	1	0	0	99	0.005	55.625				
0	0	0	1	0	1296	0.113	1296.000			$\pi^* =$	0.128
1	0	0	1	0	1132	0.099	1132.000			$R(\delta, \delta^*) =$	0.993
0	1	0	1	0	568	0.043	487.429				
1	1	0	1	0	810	0.071	810.000				
0	0	1	1	0	335	0.029	335.000				
1	0	1	1	0	662	0.049	556.696				
0	1	1	1	0	285	0.021	239.708				
1	1	1	1	0	936	0.071	814.186				
0	0	0	0	1	108	0.005	61.870				
1	0	0	0	1	86	0.005	54.041				
0	1	0	0	1	53	0.002	23.270				
1	1	0	0	1	82	0.003	38.669				
0	0	1	0	1	22	0.001	15.993				
1	0	1	0	1	52	0.002	26.576				
0	1	1	0	1	29	0.001	11.444				
1	1	1	0	1	61	0.003	38.869				
0	0	0	1	1	328	0.020	232.883				
1	0	0	1	1	387	0.034	387.000				
0	1	0	1	1	274	0.015	166.639				
1	1	0	1	1	566	0.049	566.000				
0	0	1	1	1	131	0.010	114.527				
1	0	1	1	1	389	0.034	389.000				
0	1	1	1	1	277	0.015	167.500				
1	1	1	1	1	1066	0.093	1066.000				
				Sum	13127	1.000	11444.48				

Note:
 $G^2 = 222.27$

MSPAP - π^* for 2LCA Model

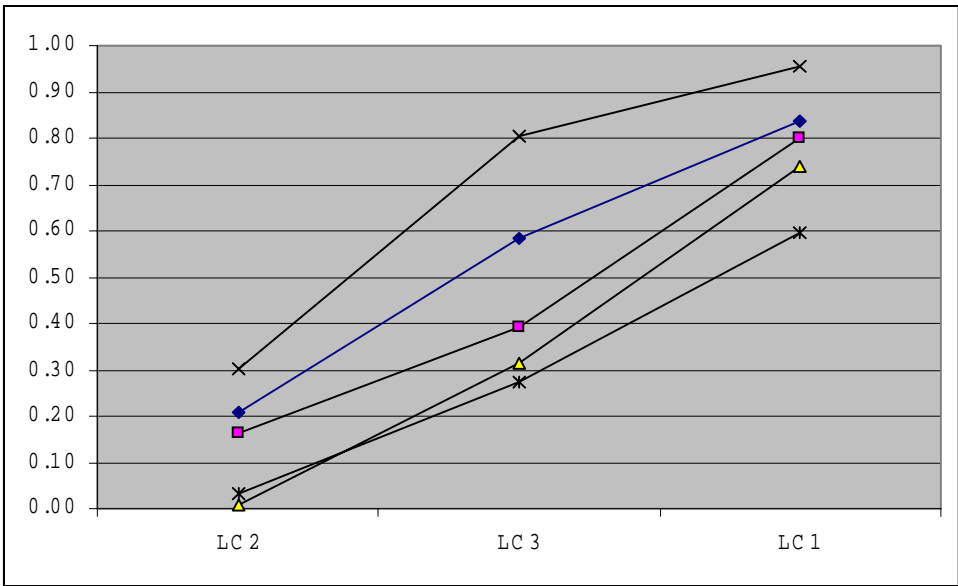
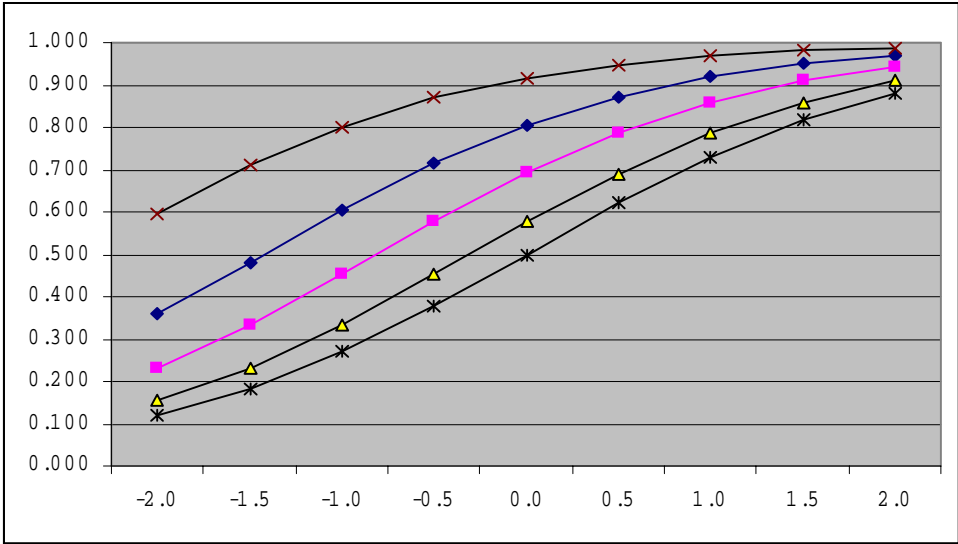
A	B	C	D	E	Freq	Prob*	E(F*)	Standard LCA		π^* Solution		
								LC1	LC2	LC1*	LC2*	
0	0	0	0	0	1614	0.141	1614.000	A	0.757	0.314	0.774	0.259
1	0	0	0	0	594	0.052	594.000	B	0.638	0.206	0.652	0.182
0	1	0	0	0	375	0.033	375.000	C	0.587	0.065	0.565	0.046
1	1	0	0	0	262	0.016	185.947	D	0.921	0.477	0.922	0.425
0	0	1	0	0	89	0.008	89.000	E	0.472	0.089	0.438	0.059
1	0	1	0	0	109	0.006	69.018	θ	0.546		0.552	
0	1	1	0	0	50	0.004	40.199	N*			11438.2	
1	1	1	0	0	99	0.007	85.085	Pi*			0.129	
0	0	0	1	0	1296	0.113	1296.000					
1	0	0	1	0	1132	0.070	797.135					
0	1	0	1	0	568	0.041	473.466					
1	1	0	1	0	810	0.071	810.000					
0	0	1	1	0	335	0.018	201.845					
1	0	1	1	0	662	0.045	516.600					
0	1	1	1	0	285	0.025	285.000					
1	1	1	1	0	936	0.082	936.000					
0	0	0	0	1	108	0.009	108.000					
1	0	0	0	1	86	0.005	60.569					
0	1	0	0	1	53	0.003	36.302					
1	1	0	0	1	82	0.005	55.410					
0	0	1	0	1	22	0.001	14.435					
1	0	1	0	1	52	0.003	34.611					
0	1	1	0	1	29	0.002	19.125					
1	1	1	0	1	61	0.005	62.122					
0	0	0	1	1	328	0.014	161.585					
1	0	0	1	1	387	0.028	324.729					
0	1	0	1	1	274	0.016	180.336					
1	1	0	1	1	566	0.049	566.000					
0	0	1	1	1	131	0.010	116.931					
1	0	1	1	1	389	0.034	389.000					
0	1	1	1	1	277	0.019	213.405					
1	1	1	1	1	1066	0.064	727.395					
			Sum		13127	1.000	11438.249					

Note:
 $G^2 = 253.31$

MSPAP - π^* for 3LCA Model

					Standard LCA			π^* Solution				
A	B	C	D	E	Freq	Prob	E(F)	LC1	LC2	LC3	LC1*	LC2*
0	0	0	0	0	1614	0.13	1614.00	A	0.836	0.209	0.582	0.798
1	0	0	0	0	594	0.05	594.00	B	0.801	0.163	0.390	0.742
0	1	0	0	0	375	0.03	375.00	C	0.741	0.010	0.314	0.709
1	1	0	0	0	262	0.02	262.00	D	0.957	0.300	0.805	0.946
0	0	1	0	0	89	0.01	89.00	E	0.594	0.035	0.273	0.549
1	0	1	0	0	109	0.01	93.36	θ	0.262	0.249	0.490	0.394
0	1	1	0	0	50	0.00	47.50	N*				12297.7
1	1	1	0	0	99	0.01	84.22	π^*				0.063
0	0	0	1	0	1296	0.11	1296.00					
1	0	0	1	0	1132	0.09	1132.00					
0	1	0	1	0	568	0.05	568.00					
1	1	0	1	0	810	0.07	810.00					
0	0	1	1	0	335	0.02	222.13					
1	0	1	1	0	662	0.04	448.83					
0	1	1	1	0	285	0.02	285.00					
1	1	1	1	0	936	0.08	935.67					
0	0	0	0	1	108	0.01	108.00					
1	0	0	0	1	86	0.01	86.00					
0	1	0	0	1	53	0.00	43.87					
1	1	0	0	1	82	0.00	59.41					
0	0	1	0	1	22	0.00	16.50					
1	0	1	0	1	52	0.00	32.21					
0	1	1	0	1	29	0.00	20.28					
1	1	1	0	1	61	0.01	65.60					
0	0	0	1	1	328	0.02	189.31					
1	0	0	1	1	387	0.02	297.66					
0	1	0	1	1	274	0.01	175.78					
1	1	0	1	1	566	0.04	500.99					
0	0	1	1	1	131	0.01	114.01					
1	0	1	1	1	389	0.03	389.00					
0	1	1	1	1	277	0.02	276.44					
1	1	1	1	1	1066	0.09	1066.00					
					13127	1.00	12297.75					

Note:
 $G^2 = 57.82$



Conclusions

The two-point mixture index of model fit proposed by Rudas, Clogg and Lindsay (1994) represents an easily interpretable descriptive measure when reporting results based on frequency data. Although computation may require some original programming, the availability of nonlinear programming algorithms greatly simplified the task. For most models, the jackknife provides a convenient and general approach to estimating a standard error for computing a lower confidence bound.

References

Clogg, C. C., Rudas, T., Xi, L. (1995) A new index of structure for the analysis of models for mobility tables and other cross-classifications. In P. Marsden (ed.) *Sociological Methodology 1995*, 197-222, Blackwell, Oxford.

Dayton, C. M. (1999). *Latent Class Scaling Analysis*. Sage Publications.

Dayton, C. M. (2002). Applications and Computational Strategies for the Two-Point Mixture Index of Fit. *British Journal of Mathematical & Statistical Psychology*, In Press.

Dayton, C. M. & Macready, G. B. (1980). A scaling model with response errors and intrinsically unscalable respondents, *Psychometrika*, 45, 343 - 356.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association*, 70, 755-768.

Rudas, T. (1999). The mixture index of fit and minimax regression. *Metrika*, 50, 163-172.

Rudas, T., Clogg, C. C. & Lindsay, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society, Series B*, 56, 623-639.

Rudas, T. & Zwick, R. (1997). Estimating the importance of differential item functioning. *Journal of Educational and Behavioral Statistics*, 22, 31-45.

Xi, L. (1994). The mixture index of fit for the independence model in contingency tables. Master of Arts paper, Department of Statistics, Pennsylvania State University.

Xi, L. & Lindsay, B. G. (1996). A note on calculating the π^* index of fit for the analysis of contingency tables. *Sociological Methods & Research*, 25, 248-259.