

INDEPENDENT, THIRD-PARTY PRINT® SURVEY EVALUATION

Dr. Rafael Alexandre dos Reis, PhD.

Data Scientist

22nd Aug 2023.

A randomly-generated dataset of 5,000 respondents to the PRINT survey was analyzed to provide evidence on the optimal number of clusters/groups that exist within the population. The 36-question initial assessment of the PRINT methodology was used. Latent Class Analysis (LCA) was used as it facilitates the identification of potential latent groups or subgroups within the data (also called clusters in this report), each representing unique patterns of responses that might signify distinct subgroups given the survey's specific domain of knowledge.

Statistical Procedure

Ten different solutions were generated in the dataset and compared using indicators of statistical quality (fit indices). The procedure started by evaluating a 2-group solution and ended up by examining a 12-group solution. Fit indices were used to evaluate the quality of each solution. In Latent Class Analysis, a fit index is used to determine how well the data fits the given grouping structure. Fit indices help in determining the optimal number of groups for the dataset, as well as in comparing the fit of different grouping solutions or models. Six different fit indices were used to perform this systematical comparison: Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), G-squared Statistic (Gsq), Chi-squared Statistic (Chisq), Entropy and Log-likelihood Difference (LLD).

Key Takeaways

- ◁ BIC and AIC suggest that a 9-cluster solution is optimal

The BIC and AIC values consistently decline as we progress from 2 to 9 clusters, suggesting that the fit of the model improves as more clusters are added up to this point.

After reaching 9 clusters, the BIC and AIC values either negligibly decrease or slightly rise, pointing towards an increase in model complexity without a corresponding benefit in fit. This indicates that the 9-cluster solution appears to be the most optimal among those assessed.

- ◁ Entropy rises between 7 and 9 clusters, indicating that these are good solutions

Entropy, which serves as a measure of the clarity or certainty of classification, rises between the solutions of 7 clusters to 9 clusters.

This increase in entropy suggests that the clarity of classification improves in this range, potentially marking this range as the most distinct and well-segregated set of clusters. A higher entropy value closer to 1 indicates clearer classification, while a value closer to 0 signifies more ambiguity in assignments to clusters.

- ◁ LLD suggests that a 9-cluster solution is best

The Log-Likelihood Difference (LLD) between subsequent cluster solutions provides insight into the improvement in model fit as more clusters are added. For example, the LLD between the 2-cluster and 3-cluster solutions is 3414, indicating a significant improvement in fit with the addition of one more cluster. However, it's noteworthy that the LLD between the 9-cluster solution and the 10-cluster solution is the lowest among all, at 347. This sharp drop in LLD suggests that the addition of a 10th cluster does not offer a substantial improvement over the 9-cluster solution.

- ◁ Chisq suggests the quality of the solutions increases relatively more until we reach the 9-cluster solution

The Chi-squared statistic (Chisq) is used to test the goodness of fit of the observed data to the expected data. For our models, there are some significant fluctuations in the Chisq values across the different cluster solutions. Particularly, there are notable spikes in Chisq values at the 3-cluster and 4-cluster solutions. However, post the 5-cluster solution, there's a consistent decrease in these values, with the lowest being observed for the 9-cluster solution. This decline further reinforces the idea that the 9-cluster model fits the observed data much better than other models, without overfitting like subsequent solutions might.

- ◁ G-squared Statistic corroborates with the previous findings

G-squared (Gsq) is another goodness-of-fit statistic, and its trend is somewhat complementary to the Chisq values. Observing the Gsq values for our models, there's a steady decline from the 2-cluster to the 9-cluster solution, showing that the model fit continuously improves as we add

more clusters up to this point. However, the rate of this decline starts to diminish after the 9-cluster solution, indicating diminishing returns in model fit improvement. This supports the notion that the 9-cluster solution strikes a balance between improving the fit and not overly complicating the model.

The following table presents the fit indices reported above.

N. Clusters	BIC Diff	AIC Diff	Gsq Diff	Chisq Diff	LLD
3	-3099	-3340	-3414	2.43E+12	3414
4	-1975	-2216	-2290	-3.06E+12	2290
5	-954	-1195	-1269	-7.17E+11	1269
6	-719	-960	-1034	-2.76E+11	1034
7	-616	-858	-931	8.90E+10	931
8	-353	-594	-669	-3.01E+11	668
9	-152	-393	-466	-1.14E+10	467
10	-31	-272	-347	-9.44E+09	347
11	-208	-449	-523	8.97E+09	522
12	-55	-296	-370	1.82E+09	371

Main Conclusion

Upon analyzing the fit indices for multiple cluster solutions, it is evident that the 9-cluster solution is optimal, thus providing robust statistical evidence that there are 9 distinct subdimensions within the examined population.