Sawtooth Software

RESEARCH PAPER SERIES

MaxDiff Analysis: Simple Counting, Individual-Level Logit, and HB

Bryan Orme,
Sawtooth Software, Inc.

***530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com***

MaxDiff Analysis: Simple Counting, Individual-Level Logit, and HB

Bryan Orme, Sawtooth Software

Introduction

In MaxDiff analysis (Louviere 1991), respondents are asked to evaluate a dozen or more items (such as attributes, statements, or brands) in sets of typically four or five at time. Within each set, they simply choose the best and worst items. Generally, an HB (hierarchical Bayes) application of multinomial logit analysis (MNL) is used to estimate individual-level scores for MaxDiff experiments. Other similar methods that develop individual-level scores by leveraging information across a sample of respondents may be employed (e.g. Mixed Logit or Latent Class with C Factors offered by Latent Gold software), and recent papers have shown comparable results among these methods (Train 2000, McCullough 2009). However, for certain situations researchers may need to compute a reasonable set of individual-level scores on-the-fly, at the time of the interview, without the ability or benefit of combining the respondent's data with a sample of other respondents.

Counts Analysis

In standard *Counts* analysis for MaxDiff, the percent of times each item is selected best or worst is computed. A simple form of summarizing MaxDiff scores combines the two measures, taking the percent of times each item was selected *best* less the percent of times each item was selected *worst*. Most researchers typically consider running counts analysis to summarize scores at the segment or population level, but this counting approach could easily be extended to the individual level given enough within-respondent information (each item evaluated enough times by respondents).

Consider a perfectly balanced design, where each item appears an equal number of times (say, 4) across the sets evaluated by a single respondent. MaxDiff questions don't allow respondents to indicate that an item is both best and worst within the same set. Thus, there are relatively few unique counting scores that can result when each item is shown 4 times to each respondent. The highest possible score occurs when the item is picked 4 times as best (100% best minus 0% worst = net score of 100). The lowest possible score is 0% best minus 100% worst = net score of -100). But, many combinations are not feasible, such as being picked 3 times as best and 2 times as worst.

The table below shows the feasible combinations (where the net score is shown for each feasible cell), assuming a perfectly balanced design wherein each item appears 4 times across the sets.

	Times Picked “Worst”					
		4	3	2	1	0
Times Picked “Best”	4					100
	3				50	75
	2			0	25	50
	1		-50	-25	0	25
	0	-100	-75	-50	-25	0

There are 9 unique feasible scores (-100, -75, -50, -25, 0, 25, 50, 75, 100) for an experiment where each item appears exactly 4 times for each respondent[1]. A typical MaxDiff dataset will investigate 20 or more items, so many of the items will receive duplicate scores for a respondent. This would seem to be a weakness, since the scale lacks the granularity offered by a truly continuous scale. But, compare this to the standard 10-point rating scales that are often used in market research for rating a series of items. With rating scales, many ties are given, and our situation with MaxDiff (which encourages discrimination), simple counting analysis, and 9 unique possible scores for each respondent seems quite adequate. Still, researchers wanting to try to identify (uniquely) the best or worst items for individual respondents may be bothered by the prevalence of ties.

Example Using Real MaxDiff Dataset

A few years ago, various Sawtooth Software users donated a few dozen MaxDiff datasets for us to use in our research. They shared just the raw data, without any labels or even revealing the product category (to protect client confidentiality). One of these datasets featured 287 respondents and 15 items in the MaxDiff study, shown in 15 sets of 5 items per set. In each set, both “best” and “worst” judgments were provided. Each respondent received a unique version of the design plan (as generated by our MaxDiff software program), where each version was nearly balanced and nearly orthogonal. Therefore, each respondent typically would have seen each item 5 times (though there are slight imbalances within each respondent’s design, so some items sometimes appear 4 times or 6 times). We used this dataset to test the quality of results of the simple counting model described above. This dataset seemed particularly good for this purpose since we could hold out 3 of the tasks for validation (the final 3 tasks), while using the other 12 tasks for computing the scores (and each item would typically have been shown 4 times across those 12 tasks, leading to a reasonable amount of information for each respondent to employ our purely individual-level analysis approach). For the validation, we used the counting scores to predict the 3 choices of best and the 3 choices of worst for the 3 holdout tasks. Therefore, there were a total of 287x6=1722 choices in our holdout validation.

We should note that even though hit rates are commonly used as a measure of predictive validity, they are somewhat insensitive for comparing the performance of different models. Models that

[1] For a study in which each item is shown 3 times per respondent, there are 7 unique scores possible from counting analysis: (-100, -67, -33, 0, 33, 67, 100). With items shown 5 times per respondent, there are 11 unique scores: (-100, -80, -60, -40, -20, 0, 20, 40, 60, 80, 100).

seem to be fairly different in their performance in terms of yielding unbiased estimates of true parameters may appear to have quite similar hit rates. Despite the weaknesses, hit rates provide an intuitive method for comparing models, so they are given here.

Comparisons to HB Analysis

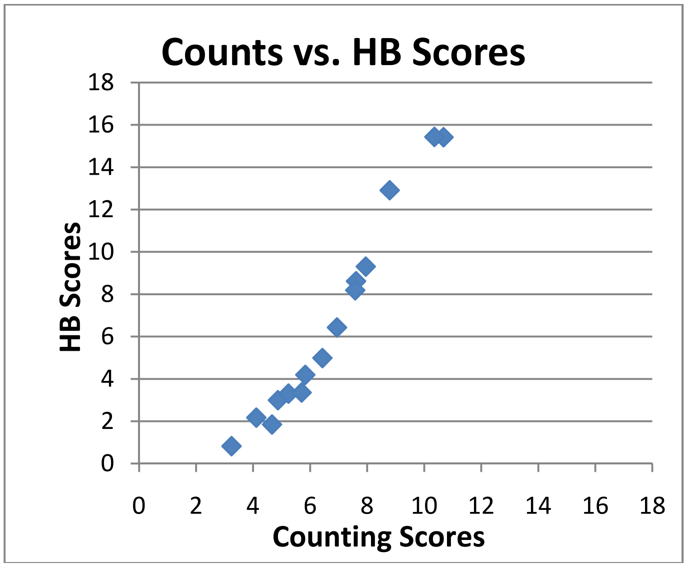
The current gold standard for estimating quality individual-level scores for MaxDiff is hierarchical Bayes (employing the MNL model), though there are a few other techniques that lead to similarly high-quality results. In each HB run, we used 10,000 burn-in iterations and 10,000 saved iterations. We used default settings for Degrees of Freedom (5) and Prior Variance (1.0). We collapsed (averaged) the saved draws into point estimates for each respondent. Using those point estimates to predict choices, we achieved a hit rate of 64.2%.

The simple form of counting analysis described earlier computes the percent of times items were picked best and worst for each respondent. We combine the two measures by simply subtracting the percent worst from the percent best for each item. We computed individual-level scores in that manner, and the resulting hit rate was surprisingly good (63.0%), given the simplicity of the approach. With 5 items per set, the hit rate due to chance is 20%, so our results are better than three times the chance level.

One concern we noted earlier is the limited number of unique scores that can result at the individual level using the counting procedure. For this dataset, there were ties for the highest-preference item for 22% of respondents (20% had the top two items tied, and 2% had the top three items tied). In cases in which tied items resulted in ambiguity regarding which item would be predicted to be chosen best or worst in the holdouts, we randomly broke the tie when computing the hit rates.

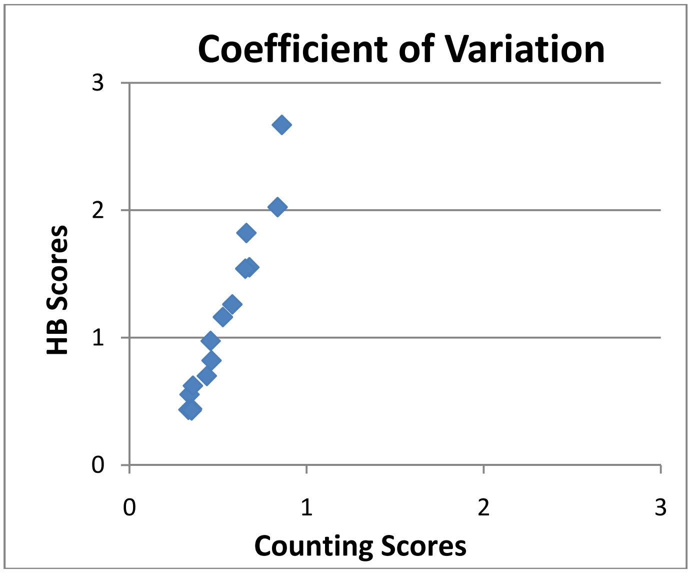
One naturally wonders how the population means and standard deviations compare between the simple counting approach and HB. This provides a way beyond simple hit rates to assess the quality of the parameter estimation (since HB is widely considered to provide unbiased parameter estimates for the population). To make the comparison, it was important to use a similar scaling for both approaches. For the counting method, the scores are computed using choice probabilities. For each respondent, for ease of presentation, we rescaled the scores so that the worst item was zero and the scores summed to 100. In the documentation of our MaxDiff software system, we describe a rescaling approach that takes raw HB scores (logit-scaled) and puts them on a scale more directly related to probabilities of choice, where the worst score for each individual is zero and the scores sum to 100. The approach involves exponentiating the raw logit-scaled parameters, leading to scores that are proportional to choice likelihood within the context of the original sets (see Sawtooth Software’s MaxDiff software documentation for more details).

A scatter-plot comparing the average rescaled Counts scores and HB scores for the sample (n=287) is given below.



Although the scales aren’t identical (HB scores show more variation), the two sets of data are highly correlated and nearly linear in their relationship. It seems that the simple counting approach produces very similar average scores on average with HB, especially if one is concerned with a rank-order prioritization of features.

Another measure to consider when comparing scores from different methods is the standard deviation across the sample (a reflection of heterogeneity). The standard deviation is naturally larger for higher-scoring items and smaller for lower-scoring items. One way to examine the variation of the score relative to its absolute magnitude is to compute a Coefficient of Variation, which is done by taking the standard deviation of the item divided by its population mean. The results are shown below, again comparing Counts vs. HB for the 15 items.



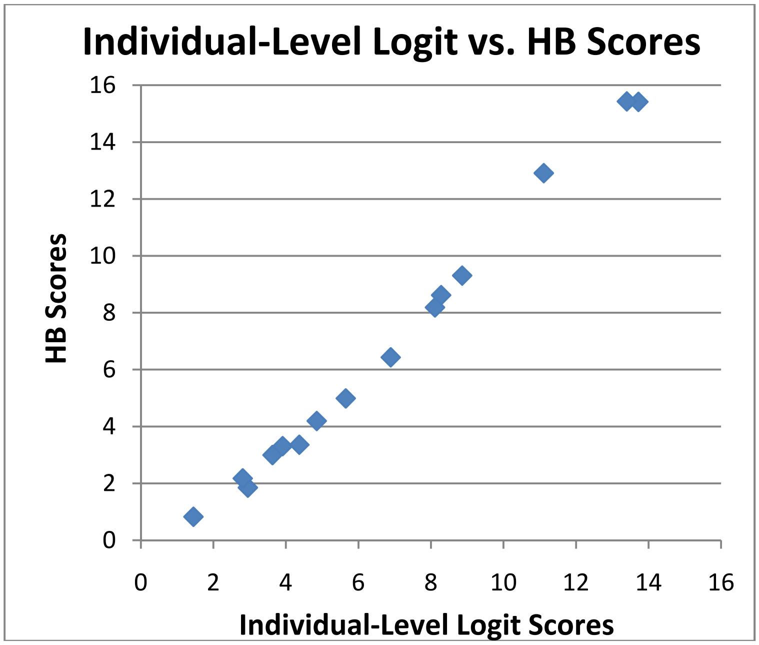
The results are highly correlated between counts and HB, so the simple method provides reasonable *relative* measures of across-sample variance for the item scores. But, the HB scores tend to have greater variation across respondents. This can be expected, since the counting method lacks granularity in its individual-level scores.

Individual-Level Logit

Multinomial Logit (MNL) has been in use since the 1970s as a parameter estimation method for choice data. MNL captures more information than simple counts because it takes into consideration not only which items were chosen, but the strength of competition within each set (which counting analysis ignores). With a perfectly balanced experiment, where for each person each item appears with each other item exactly an equal number of times, simple counts should reflect essentially full information. But, with fractional designs that are modestly imbalanced at the individual-level, MNL is a more complete and robust model for estimating strength of importance/preference for the scores in a MaxDiff experiment.

Using individual-level MNL[2], we achieved a hit rate of 63.3%, only slightly better than Counts (63.0%) and a little bit lower than HB (64.2%). Naturally, this leads one to wonder whether the more sophisticated method was worth the effort. Again, hit rates are a blunt instrument for gauging the quality of models, and it is helpful to look at other measures, such as the recovery of parameters.

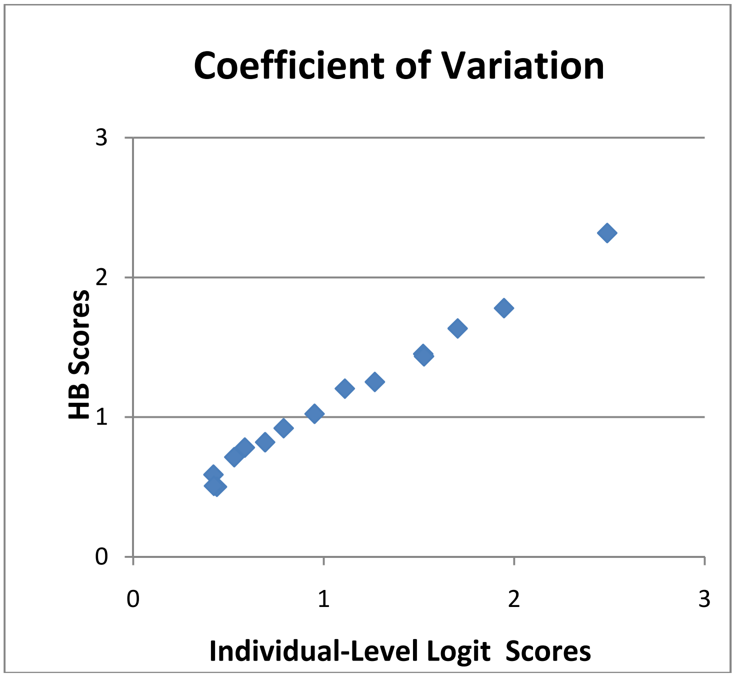
We previously compared the mean scores for the sample (n=287) for Counts vs. HB and found close congruence. The same chart for individual-level logit scores vs. HB is given below.



[2] In our MNL implementation, we used the Counts solution described above as the starting point, and we iterated until a stopping point criterion of the log-likelihood not improving by more than 0.2 for each individual. The computations are extremely quick for each individual, solving in less than a second.

There is strong correlation between the two sets of scores (even stronger than counts), with the values falling nearly on the 45-degree line.

The chart below compares the coefficient of variation between individual-level MNL and HB.



Whereas counting analysis didn't demonstrate as much absolute discrimination across respondents as HB scores, individual-level MNL scores reflect very similar absolute measures of variation for the items as HB.

With our MaxDiff dataset, there is so much information at the individual level that our HB version of MNL employs only a modest amount of Bayesian shrinkage toward the population parameters. Thus, it's no surprise that the purely individual-level MNL results seem so similar to the HB solution.

Conclusions

If one has enough information at the individual level in MaxDiff studies (e.g. each item shown about 4x per respondent), purely individual-level analysis is feasible. HB (or related methods) provides a more robust overall solution, as it can leverage both the information at the individual-level and data across other respondents within the sample. But, if one needs purely individual-level scores on-the-fly, individual-level MNL provides very similar results as HB given the quantity of individual-level information for our dataset. Even the simple method of Counts seems to provide reasonable estimates for this dataset, where the designs are quite balanced and nearly orthogonal.

References:

Train, Kenneth (2000), “Estimating the Part-Worths of Individual Customers: A Flexible New Approach,” Sawtooth Software Conference, Sequim, WA.

Louviere, J. J. (1991), “Best-Worst Scaling: A Model for the Largest Difference Judgments,” Working Paper, University of Alberta.

McCullough, Paul (2009), “Comparing Hierarchical Bayes and Latent Class Choice: Practical Issues for Sparse Datasets,” Sawtooth Software Conference Proceedings, Forthcoming, Sequim, WA.