Sawtooth Software

RESEARCH PAPER SERIES

Shapley Values for MaxDiff Results: Overview and Recommendations for Practitioners

Bryan Orme
Sawtooth Software
David W. Lyon
Aurora Market Modeling LLC

3210 N. Canyon Rd., Provo, Utah
+1 801 477 4700
www.sawtoothsoftware.com

Shapley Values for MaxDiff Results: Overview and Recommendations for Practitioners

Bryan Orme (Sawtooth Software) and David W. Lyon (Aurora Market Modeling)

Principal Motivation

Not long after Sawtooth first offered TURF (Total Unduplicated Reach and Frequency) as an optimization extension for MaxDiff analysis, users said clients wanted to know the answer to this key question: "For a portfolio of four items with 80% reach, how much does each of the four items contribute to its reach?"

The answer to that question depends on the order of entry we assume for those items into the four-item portfolio. To fairly account for how much each item contributes to total reach, we need to consider the average marginal contribution of each item across all possible subsets and orderings. And, that's what Shapley values (an idea originated in cooperative game theory) do.

(Note: Shapley values can be applied to all sorts of data, including pick-any data and rating scales. This article focuses on its application to TURF analysis of MaxDiff data.)

Is TURF Any Different from Average MaxDiff Scores?

The motivation behind TURF is that the optimal four items to include in a portfolio aren't necessarily the four best items in terms of average utility scores. High average utility items could have redundancies in terms of motivating the same groups of people. The classic examples involve magazines to advertise in to reach the most unique readers, or ice cream flavors to stock on the shelf such that the maximum number of customers find at least one flavor they'd want to eat.

The top four magazines in terms of readership may not maximize the total number of eyeballs (the reach) for your ad, because there can be significant overlap in readership on certain highly popular magazines. Some magazine titles might fall in 8th or 9th place overall in terms of overall readership, but make it into the optimal top set of four magazines to advertise in because they are read by a large enough group of people that are not reached by the other three. A similar argument can be made for ice cream flavors.

Further, if we're interested in quantifying the contribution of each of four items within an optimal (or near-optimal) portfolio, it's not as simple as saying the contributions are proportional to their average utility scores. Some items can punch higher than their weight (using a boxing analogy) in terms of avoiding redundancies and reaching more unique consumers. These same types of items also get promoted in "importance" for Shapley values that are computed based on TURF reach.

Extending the Idea

Rather than just limiting ourselves to the specific case of, say, a 4-item portfolio, we can extend the Shapley Value calculation across a range of portfolio sizes (e.g., 3 to 6 items) or even the entire set of items in a MaxDiff study. For a 20-item study, we can compute TURF reach for all portfolios of size 1 through 20, considering all possible orders of entry for the items. Shapley values summarize the *average* contribution of each item to total reach, averaged across all possible inclusion orders. They bridge the gap between portfolio optimization and a more generalized explanation of which items are adding to the reach of groups or portfolios of items. In many situations, TURF included, Shapley values can be thought of as the average reach improvement when an item is added to any combination that does not already include it.

But, There's a Catch!

But there's a problem: the number of permutations expands factorially. A 20-item MaxDiff involves 20! permutations to compute Shapley values across the entire set of items. That's 2.4 quintillion combinations—too many to compute in reasonable time. As a workaround, researchers have approximated Shapley values by sampling across multiple thousands of randomly drawn order of entry permutations.

At the 2026 Sawtooth Research Conference, Dave Lyon described "Component Order of Addition" or "COA" designs. They are compact order of entry designs that are more accurate and efficient than random sampling for computing Shapley values, and workably small even for very large Shapley problems. These designs make it practical, accurate, and efficient to compute Shapley values for TURF problems involving, say, 20, 100, or more MaxDiff items.

Enhanced Stability

A challenge with TURF is that small changes to the dataset (such as deleting a few respondents or adding a few more) can lead to changes in the composition of the optimal few portfolios. Minor underlying changes between subgroups or between tracking waves can cause sudden shifts in TURF results. In contrast, Shapley values are more robust to small changes in the dataset. That stability is important for managerial decision-making. (See more on this in Table 1 below.)

Given this advantage, Shapley values also help answer another common client request. Clients often ask for the buildup involving the optimal 2-item, 3-item, 4-item, etc. portfolios. What often happens is that the optimal k+1 size portfolio doesn't include the original k items—another item can switch in, which is entirely accurate but can be unsettling to the client. Shapley values give an appropriate and stable result for reporting the buildup of items into optimal portfolios of increasing size. To report the top 3-item portfolio, just take the top 3 items in terms of Shapley values. To report the top 4-item portfolio, take the top 4 items on the Shapley values, etc. Those top few items are not guaranteed to be the *absolute* best combination of a given size, but they will at least be very close, and will give a result that does not over-react to minor sampling changes.

What Can Shapley Values Do that Average HB Scores Don't?

Shapley values assign higher weight to those items that are both preferred and uniquely reach subgroups of respondents who aren't as easily reached by more popular items. If your prioritization opportunity could benefit from that emphasis, then you'd rather be looking at Shapley values for the 20 items than average utilities. And, if your goal was to find the optimal set of, say, four items that reach the largest number of people, then a targeted TURF optimization will do that.

Which Reach Method to Use?

A TURF analysis requires us to first define what it means to reach a respondent. Is the respondent reached only if their top item is in the portfolio ("first choice reach")? Are they reached if at least one item exceeds the anchor utility ("threshold reach")? Are they at least partially reached by every item in a portfolio in proportion to the underlying logit-scaled utilities ("weighted by probability reach")? Those are reach criteria supported in Sawtooth, but there are many other potential reach definitions.

Shapley values only provide new insight when the chosen reach definition allows items to differentiate based on non-redundant audience coverage—not just overall preference.

TURF using first-choice reach assumes that each respondent is only influenced by their top item. For the optimal 4-item portfolio, it leads to simply grabbing the 4 items that are the topmost items for the greatest number of people. Shapley values for TURF with first choice reach also lead to that same limiting solution; the Shapley values simply re-iterate the first-choice top winners.

Another mostly redundant Shapley result is found when using Sawtooth's "weighted by probability" reach criterion. Those results are typically correlated about 0.98 or better with the average of the rescaled by probability HB scores (the positive values summing to 100).

To leverage the Shapley approach to reward items that are both preferred by a lot of people and uniquely reach certain consumer segments, you should specify a reach criterion that is different from first choice or weighted by probability.

There are many reach methods that can lead to different insights and different prioritization of items than average rescaled by probability HB scores. Two such approaches available in Sawtooth are:

1. **Threshold reach**: if a portfolio includes an item that exceeds a certain utility threshold, such as beating the anchor in anchored MaxDiff, then the respondent is reached by that item and by any portfolio including that item.
2. **Weighted by probability with post-thresholding**: if the weighted by probability reach for a combination or portfolio of items exceeds some threshold, such as 90%, then the respondent is reached. Consider three items with raw HB scores of i, j, and k. The portfolio's weighted probability of reach for a standard MaxDiff study is: $P_{ijk} = (e^{Ui} + e^{Uj} + e^{Uk}) / (e^{Ui} + e^{Uj} + e^{Uk} + a - 1)$, where the utilities are assumed to be zero-centered and *a* is the number of items shown to respondents in each MaxDiff set. For an anchored MaxDiff, the weighted probability of reach is computed as $P_{ijk} = (e^{Ui} + e^{Uj} + e^{Uk}) / (e^{Ui} + e^{Uj} + e^{Uk} + 1)$, where the utilities have been scaled such

that the anchor utility is 0. For this latter case, perhaps a 50% threshold should be used. (Notes: weighted by probability equations were suggested by Kenneth Fairchild. Applying post-thresholding to the weighted by probability as a reach criterion was suggested by Dave Lyon in his 2026 Sawtooth Research Conference presentation.)

Split-sample reliability simulations using a real MaxDiff dataset with about 250 respondents show that Weighted by Probability with post thresholding reach produces more stable (reliable) Shapley results than threshold reach or first-choice reach (Table 1).

Table 1: Reliability (Stability) of Different Shapley Reach Criteria

Shapley Reach Method	**Reliability (split-sample correlation)**
Post-Thresholding	0.929
Thresholding (pre-thresholding)	0.881
First Choice	0.879

(Results in Table 1 should only be interpreted as relative reliability scores, as we would probably have achieved overall much higher split sample correlations with a larger sample size dataset.)

Are Shapley Values Adding Value?

The easiest way to check that Shapley values are adding any new information (such as penalizing for redundancies) for your MaxDiff study is to scatter-plot the average rescaled by probability HB scores on the X axis against the Shapley values on the Y axis. If the points fall tightly on a straight line or smooth curve, the Shapley values are not giving any different prioritization of the items than the average utility scores give. But, when some items show a strong departure from the line or curve, the Shapley values are adding a fresh insight: those items are either punching above or below their weight in terms of reaching unique groups of respondents who are not reached with more generically preferred items. Items with much higher average utility but lower Shapley values are items with higher redundancy—they tend to reach only respondents easily reached by other items.

Managerial Importance

Shapley values as a natural extension of TURF equip managers to go beyond identifying specific optimal portfolios, allowing them to more generally understand and defend the unique role each item plays in expanding market reach. This leads to more confident, stable, and strategically grounded decisions—especially when balancing broad (but potentially redundant) appeal against the need to capture otherwise overlooked or underserved segments.

Shapley values can also simplify reporting and provide sharper insights because they are one score per item, rather than multiple TURF solutions for portfolios of varying sizes.