Sawtooth Software

RESEARCH PAPER SERIES

Rating Scales for Use in Driver Analysis

Keith Chrzan
Sawtooth Software, Inc.

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

1.0 Background on Driver Analysis

When designing a battery of questions to support driver analysis, we as researchers need to decide which scales to use for two kinds of questions. First, we need to get the respondents to provide an overall measure for the brand, service, experience, etc. we want to study. This overall measure, the measure we want to build a model to predict, we call the dependent variable. In addition, we need to elicit responses about the performance of the brand (service, experience, etc.) with respect to each of several (usually 10 – 20) attributes. These attributes (aspects of the brand) are the predictors or the independent variables in our driver analysis model.

For example, if measuring customer satisfaction with casual dining restaurants, we might ask for the respondent's overall satisfaction with their dining experience and then about the performance of the restaurant on attributes like these:

- Food served at the proper temperature
- Attentive server
- Appropriate pacing of the meal
- Good tasting food
- Reasonable price
- Restaurant cleanliness
- Server friendliness
- Good selection of meal choices
- Etc.

We then use the attributes/independent variables to predict the overall/dependent variable and we get a set of "derived importances," the importance weights that emerge from our statistical modeling. For details and a comparison of the models that produce these weights please see our white paper on driver analysis here (https://sawtoothsoftware.com/resources/technical-papers/driver-analysis); long story short, the collinearity that's present in nearly all attribute rating scales in survey research invalidates both regression and correlation analysis as engines for driver analysis and we need to go with a more robust approach like the relative weights analysis used by the Sawtooth Software driver analysis program.

Notice that we suggest putting the overall question first and the attributes after. While the opposite (attributes first, overall second) tends to result in higher R-squared statistics (implying more of the variation in the overall measure is explained by the attributes) we think this is a survey artifact – that by focusing respondents' attention on the attributes, they are to a greater extent focusing ONLY on those attributes in forming their overall impression. Moreover, if the overall is influenced by the attributes, changes to the attribute battery could cause spurious changes to the overall rating, which can be injurious in a tracking environment. As a result, we recommend getting a clean measure of the overall measure that's not influenced by the context effect of seeing the attributes.

2.0 Background on Rating Scales

The science of measuring respondents' mental states has been called psychometrics or psychophysics. Hipparchus, a Greek mathematician, is credited with the first use of a rating scale, a 6-point scale for measuring the brightness of stars. The earliest reference I can find of anyone using rating scales for psychological entities dates to 1692 when Christian Tomasius, a professor at the University of Halle, in Germany, used a 12-point scale to measure properties of individuals (seriousness, ambition, etc.).

Rating scales have become the primary tool marketing researchers use to get respondents to report the strength, degree or intensity of beliefs and perceptions.

The plethora of published research on the topic of scales gives the appearance that we should know a lot more about rating scales than we do. Much of the literature has to do with Likert scales in public policy research, so it may not be directly relevant to what we do in the world of marketing research. Much of the rest of the literature is inconsistent and offers little support to general rules about things like scale balance and number of scale points, or about which adjectives we should use as modifiers to verbal anchors. Contentious and often shrill debate occurs among scale fetishists who base strong beliefs on weak and inconsistent evidence.

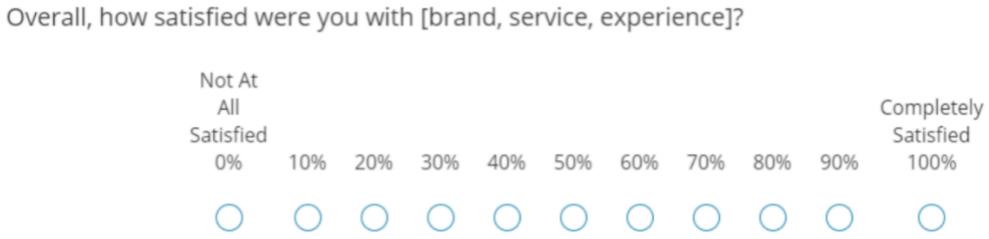
3.0 Sawtooth Software Recommendations on Rating Scales for Use in Driver Analysis

Many uses of driver analysis occur in the context of tracking studies – here the value of continuity usually outweighs the value of changing to better scales, even when better scales are known to exist. Similarly, many companies that perform concept testing surveys have normative databases of past concept tests, and again the value of continuity usually outweighs any value that could be gained by changing scales.

For applications where we want to perform a driver analysis and the above cases for continuity do not apply, Sawtooth Software recommends the following scales for driver analysis. If you want to know the evidence and rationale for these choices, check out Section 4.0 below.

Overall satisfaction:

In the empirical test described below, the top performing satisfaction scale is this one:



The runner-up, a scale that MAY work better in cross-cultural satisfaction research and which respondents enjoyed using more than the numerical scale above, uses happy, neutral and sad face icons for scale points and you can find it in the Appendix.

Overall loyalty:

For infrequently purchased product categories or continuing services we recommend:

How likely are you to recommend Acme widgets to a friend or colleague?

Definitely would NOT recommend	Probably would NOT recommend	Might or might not recommend	Probably would recommend	Definitely would recommend
○	○	○	○	○

For frequently purchased products we recommend:

Think of the next 10 times you need to buy widgets. How many of those times do you think you will buy Acme widgets?

[ ] Times

<u>Purchase intent scales for concept testing and brand image research</u>:

Many commercial forecasts use this 5-point purchase intention scale:

Taking everything into account, how likely are you to buy Acme Brand next time you need a widget?

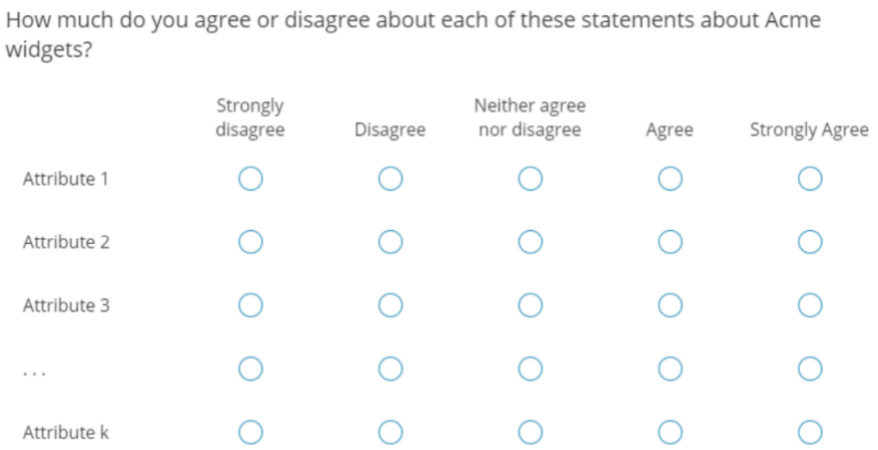
- ○ Definitely would NOT buy
- ○ Probably would NOT buy
- ○ Might or might not buy
- ○ Probably would buy
- ○ Definitely would buy

Alternatively, you can also use a Juster scale (Juster 1966):

Taking everything into account, how likely are you to buy Acme Brand next time you need a widget?

- ○ No chance, almost no chance (1 chance in 100)
- ○ Very slight possibility (1 chance in 10)
- ○ Slight possibility (2 chances in 10)
- ○ Some possibility (3 chances in 10)
- ○ Fair possibility (4 chances in 1\0)
- ○ Fairly good possibility (5 chances in 10)
- ○ Good possibility (6 chances in 10)
- ○ Probable (7 chances in 10)
- ○ Very probable (8 chances in 10)
- ○ Almost sure (9 chances in 10)
- ○ Certain, practically certain (99 chances in 100)

Attribute ratings for all the above:



4.0 Empirical Evidence and Rationale for Recommended Rating Scales for Use in Driver Analysis

Prior to joining Sawtooth Software, I was Chief Research Officer at Maritz Research. There I invested my R&D budget heavily on a series of tests of rating scales for use in customer satisfaction, loyalty, and brand research, not coincidentally three of the most common places researchers run driver analyses (the fourth being new product concept testing research). Pulling together scales suggested in the academic literature we subjected them to various tests of reliability (test-retest reliability) and validity (specifically of criterion, concurrent and predictive validity). For the customer satisfaction and loyalty scale tests we studied 5 product categories (automotive, video game rentals, retail banking, mobile phones and hotels) and we conducted surveys among over 3,200 respondents. We even tried to re-interview respondents 4 weeks after their initial surveys and we managed to complete these recontact surveys with 64% of our initial respondents. These recontact surveys allowed us to assess predictive validity and test-retest reliability.

While we identified clear winners and losers in the customer satisfaction and loyalty spaces, we never published the results because I left the company and then company opted to forego marketing research and become a customer experience software company. The suggestions below reflect the results of these rigorous empirical tests. Also, below you'll find scale recommendations based on less clear-cut test results for brand research and we extend the findings from customer satisfaction, loyalty and brand research to make recommendations for concept testing research.

4.1 Overall Satisfaction

In our large tests over six product categories we tested and compared nine different overall rating measures (see Appendix for examples):

- A 5-point unipolar scale with verbal anchors on each scale point, adapted from Aiello and Czepiel (1979)

- A 5-point bipolar scale with verbal anchors on each scale point (used by many of Maritz's clients)
- A 7-point terrible-to delighted scale with verbal anchors on each scale point and two off-scale points for "neutral" and "I never thought about it," borrowed from the academic literature on life satisfaction by Andrews and Withey (1974)
- A 9-point bipolar scale using Trivial Pursuit-style pie slices with endpoints and midpoint anchored, suggested by Westbrook (2012)
- A 7-point scale with images ranging from a happy face to a sad face and no verbal anchors
- An 11-point percentage scale akin to a Juster scale of purchase probability (Westbrook 1980)
- A binary scale of satisfied or not clients
- A 7-point bipolar expectations scale with endpoints and midpoint labeled

To avoid going through the results of each of our various tests in excruciating detail, the chart below summarizes our findings with respect to (roughly in decreasing order of importance):

- Test-retest reliability - how stable is the score when measured at twice, four weeks apart?
- Convergent validity – how correlated is the measure with the other 8 measures of satisfaction?
- Criterion validity A – how well is the measure predicted by drivers?
- Criterion validity B – how well does the measure predict attitudinal loyalty?
- Respondent evaluation – how much do respondents like the scale?

Scale	Test-retest reliability	Convergent Validity	Predictions by drivers	Prediction of loyalty	Predictive validity	Respondent liking
Pie slices	☺				na	
5 point unipolar				☺	☺	
Delighted-terrible		☺		☺	☺	
5-point bipolar					na	☺
Expectations					na	
Faces	☺	☺	☺	👍	na	👍
11-point percentage	👍	👍	👍		☺	☺
Binary	☠	☠	☠	☠	☠	☠

👍 - Better than all other measures

☺ - Better than some other measures

☠ - Worse than all other measures

na - Not measured

The binary satisfied-not scale performed significantly worse than the other scales on every measure. The expectations scale scored better than the binary scale but worse than all the other options.

The 11-point percentage scale performed best followed by the 7-point non-verbal faces scale. The 11=-point scale was also the most able to discriminate better from worse brands.

The two 5-point scales, the delighted-to-terrible scale and the pie faces scales filled out the middle of the pack.

4.2 Satisfaction Attribute Ratings

In terms of rating scales for the attributes, two scales are by far the most common:

- A 5-point fully anchored bipolar agreement scale (Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree)
- A 5-point fully anchored bipolar performance scale (excellent, very good, good, fair, poor)

We did not test, nor have I seen tests, of using one as opposed to the other in customer satisfaction research. I prefer the former, a commonly used scale for attitudinal research and in my personal experience it works well.

4.3 Overall Loyalty

In one loyalty study, a 5-point intent to recommend scale (definitely would, probably would, might or might not, probably would not, definitely would not) outperformed other loyalty measures in terms of both test-retest reliability and in terms of convergent validity. In the other study of loyalty in a frequently purchased product category, we tested intent to recommend, intent to return and a probability allocation ("think of the next 10 times you will [X]. How many of those times do you think you will use <BRAND>"). The allocation measure had the best predictive validity (for subsequent purchase behavior) and the greatest ability to discriminate among brands. As a result, we suggest the probability allocation for frequently purchased product categories and the intent to recommend measure for infrequently purchased product categories.

4.4 Loyalty Attribute Ratings

As for customer satisfaction, I have seen no published evidence of the superiority of performance scales over agree-disagree scales or vice versa, so I tend to prefer the agree-disagree scales.

4.5 Concept Testing Overall Purchase Intent Rating

We recommend one of two measures for purchase intention, both of which can be found in the bible of new product research by Urban and Hauser (1993):

- A 5-point verbally anchored bipolar purchase intention scale (definitely would buy, probably would buy, might or might not buy, probably would not buy, definitely would not buy). Used by many FMCG firms, this purchase intention scale is a good candidate, particularly if the client has developed norms for the purchase intention ratings.
- In other industries, however, the 11-point Juster scale, shown in Appendix 2, seems to work just fine as well. If one wants to standardize the way they measure in customer satisfaction and brand research, the similarity of the Juster scale and the 11-point satisfaction scale may make those the ones to standardize on.

4.6 Concept Testing Attribute Ratings

Again agree-disagree scales and performance scales both seem to work fine for driver analysis of concept test results, and I'm unaware of research proving one to be superior to the other.

4.7 Overall Brand Rating

In brand research respondents typically rate multiple brands and it becomes very important to use discriminating measures that can distinguish better from worse brands.

While some brand studies use brand liking as the overall measure, respondents can like brands they never intent to buy. For this reason, it makes more sense to use an overall measure that relates to purchase likelihood, using the concept testing purchase intention scales above.

Sawtooth Software users will realize that brand studies featuring ratings of multiple brands are also amenable to modeling via the multinomial logit (MNL) choice model instead of regression-based driver analysis. Recommendations for using MNL and other choice models for brand research can be found in Chrzan and Malcom (2009). When using this model, it can be important to account for the halo effect, and for this we like the double-centering approach of Dillon, Mulani and Frederick (1984). Assuming you don't go down this choice modeling path, and that you stick with driver analysis, most of my clients find it useful to run drivers for individual brands AND on a "stacked" data set that combines all the brands into one driver analysis.

4.8 Brand-Attribute Ratings

With respondents rating multiple brands on multiple attributes, the brand researcher may want to use scales that make life easier for respondents. As reported in the 2007 Sawtooth Software Conference Proceedings, Doug Malcom and I compared rating scale measures to more respondent-friendly binary (yes/no and pick-any) measures (Chrzan and Malcom 2007). Unfortunately, the binary scales ended up providing little signal beyond the halo effect, so they did not seem to be valuable replacement for rating scale measures. While Doug and I tested semantic differential ratings and comparative ratings, we see a lot of brand studies using agree-disagree ratings and performance ratings and I'm not aware of evidence that supports any one of these more than the others (again at least among scales with more than two points).

References

Aiello, A. and J.A. Czepiel (1979) “Customer Satisfaction in a Catalog Type Retail Outlet: Exploring the Effect of Product, Price and Attributes,” in *New Dimensions in Consumer Satisfaction and Complaining Behavior*, eds. R.L. Day and H.K. Hunt, Bloomington: Indiana University, 29-135.

Andrews, F.M. and S.B. Withey (1974) "Developing Measures of Perceived Life Quality," *Social Indicators Research*, **1**, 1-26.

Chrzan, K. and D. Malcom (2007) “An Empirical Test of Alternative Brand Measurement Systems,” *Sawtooth Software Conference Proceedings*, Sequim: Sawtooth Software, 155-167.

Chrzan, K. and D. Malcom (2009) “How to Improve Brand Tracking Research: A Frozen Pizza Case Study,” *International Journal of Market Research*, **51**, 723-733.

Dillon, W.R., N. Mulani and D. Frederick (1984) “Removing Perceptual Bias in Product Space Analysis,” *Journal of Marketing Research*, **21**, 184-193.

Juster, F.T. (1966) “Consumer Buying Intentions and Purchase Probability: An Experiment in Survey Design,” *Journal of American Statistical Association*, **61**, 658-696.

Urban, G.L., and J.R. Hauser (1993) *Design and Marketing of New Products, 2nd ed*. Englewood Cliffs: Prentice-Hall.

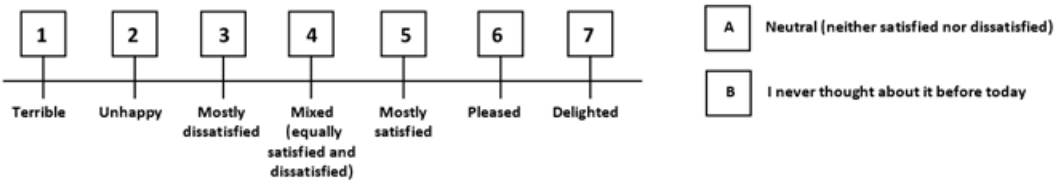
Westbrook, R.A. (1980) “A Rating Scale for Measuring Product/Service Satisfaction,” *Journal of Marketing*, **44**, 68-72.

Appendix – Overall Satisfaction Scales Tested

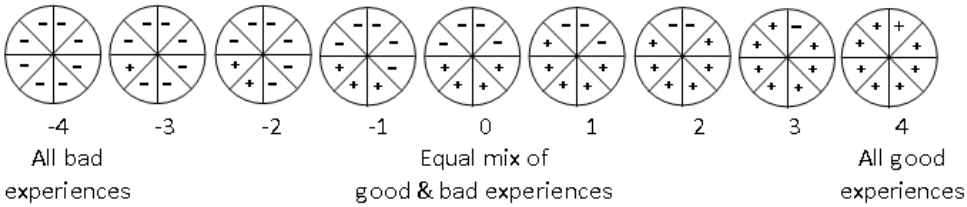
Five-point unipolar satisfaction scale:

Five-point bipolar satisfaction scale:

Delighted-Terrible scale:



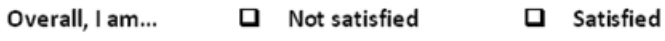
Pie slices scale:



Faces scale:

11-point percentage satisfaction scale:

Binary satisfaction scale:



Expectations scale: