Sawtooth Software

RESEARCH PAPER SERIES

Number of Levels Effect in CBC: Is It Strong and Does It Persist for More than Four Levels?

Bryan Orme
Zachariah Hewett
Sawtooth Software, Inc.

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

Number of Levels Effect in CBC: Is It Strong and Does It Persist for More than Four Levels?

Bryan Orme, Sawtooth Software

Zachariah Hewett, Sawtooth Software

Executive Summary:

Researchers in the 1980s and 1990s found a number of levels effect in conjoint analysis when specifying a quantitative attribute on four versus two levels (holding the range of variation constant). An attribute such as price specified on four levels rather than two levels would capture significantly more importance, as indicated by the difference in utility between best and worst levels. We report on two CBC studies on the number of levels effect. Our research suggests that the number of levels effect in conjoint analysis (Choice-Based Conjoint) is not as strong as previously thought. Additionally, we didn't observe any increase in attribute importance moving from 11 to 21 levels of a quantitative attribute. When using quantitative attributes in conjoint analysis, we recommend using at least four levels to both reduce the number of levels effect and represent these attributes in a way that is likely more realistic.

Background on the Number of Levels Effect:

In the 1980s, academics (Currim et al. 1981) showed that by doubling the number of levels from two to four for a quantitative attribute, the importance of that attribute would increase significantly (holding the range of variation constant). For example, some respondents would see a two-level conjoint attribute with prices of $100 and $200. Other respondents would see price defined on four levels ($100, $130, $160, $200) and the difference in utility between $100 and $200 would be significantly larger for the second group compared to the first group. How much larger? In one case, Wittink reported for full-profile ratings-based conjoint that increasing the number of levels for an Energy Cost for refrigerators attribute from two levels to four levels (holding the range of variation in Energy Cost constant) led to a more than doubling of the importance of the attribute (Wittink et al. 1991).

A key aim of our investigation was to examine if the number of levels effect for conjoint analysis continues for numbers of levels for quantitative attributes beyond four. Since the 1980s, we're only aware of one other piece of research that examined the number of levels effect in conjoint analysis for numbers of levels beyond either two or four. We wrote about that research (by Marco Hoogerbrugge) in our "Becoming an Expert in Conjoint Analysis" book:

> In 2000, Marco Hoogerbrugge conducted an experiment in which respondents got one of multiple CBC questionnaires that varied the number of levels of quantitative attributes while holding their ranges constant (Hoogerbrugge 2000). He tested attributes with three, five, eight, and nine levels (holding their ranges constant). Hoogerbrugge found that the number of levels effect seemed muted and potentially non-existent when comparing three or more levels to a larger number of levels such as five, eight or nine. Since most conjoint researchers don't consider using just two levels of quantitative attributes, but rather

> consider more levels such as three to five, it seems like the number of levels concern is much less of a problem than it was first considered to be.
>
> The number of levels effect may also be present in the real world for qualitative attributes like brand and color. The more brands or colors available on the shelf, the more the buyer's attention may be drawn to those product attributes. We suggest trying to mimic real world choices in CBC experiments, so if the number of levels of one attribute is much greater than another in the real world, it would seem proper to reflect that in our conjoint designs as well.
>
> --(Orme & Chrzan, 2021)

Data Collection:

Between January and March 2023, we fielded two 5-7 minute surveys. We conducted one survey using PureSpectrum[1] panel (n=1013, after cleaning) and the other using Amazon's Mechanical Turk panel (n=1400, after cleaning). Although we realized Mechanical Turk sample was not of high quality and not representative, we supposed it would be suitable for the purposes of testing the number of levels effect (after cleaning bad respondents). The subject matter was HDTV and electric vehicle preferences for MTurk and PureSpectrum respondents, respectively.

We fielded both surveys using Sawtooth Software's Discover platform. The surveys included a MaxDiff section (six items, showing six questions each with three items), five different CBC questionnaires (each respondent was randomly assigned to one of the five CBC questionnaires), and a few other opinion and demographic questions. The CBC design featured eight CBC tasks on three attributes showing three concepts per task plus a None alternative (see Appendix A).

For one quantitative attribute in each CBC survey (Price for HDTVs and Range for electric vehicles), we varied the number of levels between the experimental cells of our design.

Exhibit 1

HDTV Attributes and Levels

Resolution (2 levels): 4K, 8K

Screen Size (3 levels): 55-inch, 65-inch, 75-inch

Price: (varied for 5 different experimental design cells)

Cell 1: $500, $1000

Cell 2: $500, $750, $1000

Cell 3: $500, $650, $800, $1000

[1] Many thanks to PureSpectrum for donating the sample for this research-on-research study.

Cell 4: $500, $550, $600, $650, $700, $750, $800, $850, $900, $950, $1000

Cell 5: $500, $525, $550, $575, $600, $625, $650, $675, $700, $725, $750, $775, $800, $825, $850, $875, $900, $925, $950, $975, $1000

EV Attributes and Levels

Brand (3 levels): Tesla, Ford, Toyota

Vehicle Type (4 levels): 2-door, 4-door, SUV, Pickup Truck

Driving Range per Charge: (varied for 5 different experimental design cells)

Cell 1: 150 miles, 450 miles

Cell 2: 150 miles, 300 miles, 450 miles

Cell 3: 150 miles, 250 miles, 350 miles, 450 miles

Cell 4: 150 miles, 180 miles, 210 miles, 240 miles, 270 miles, 300 miles, 330 miles, 360 miles, 390 miles, 420 miles, 450 miles

Cell 5: 150 miles, 165 miles, 180 miles, 195 miles, 210 miles, 225 miles, 240 miles, 255 miles, 270 miles, 285 miles, 300 miles, 315 miles, 330 miles, 345 miles, 360 miles, 375 miles, 390 miles, 405 miles, 420 miles, 435 miles, 450 miles

For both the HDTV and EV projects, we included a short MaxDiff exercise to catch random responders and to prime/warm up respondents to provide better CBC data on HDTV or EV purchase choices. The MaxDiff included six items regarding buying habits of technology products for the home (for the HDTV study) and for the EV study the MaxDiff items involved six features of electric vehicles. For both MaxDiff exercises, we showed three items per task, so each item appeared three times per respondent. Using 300 random responders (bots generated in Sawtooth Software's Lighthouse Studio), the 80% HB RLH cutoff to identify random responders (per recommendations from Chrzan/Orme) was 0.431 (chance level is 0.333). 59% of MTurkers failed the consistency check from MaxDiff and were hardly distinguishable from random responders. This seems higher than usual compared to other panels in the market research industry. Only 21% of PureSpectrum respondents failed the consistency test.

Sample sizes per cell (after cleaning the random-looking respondents) were:

Exhibit 2

	MTurk HDTV	PureSpectrum EV
Cell 1	110	171
Cell 2	108	162
Cell 3	116	162
Cell 4	115	148
Cell 5	126	158
Total	575	801

Number of Levels Effect Findings:

The aim of this research was to measure what happens with the number of levels effect as the number of levels increases beyond four. Academics in the 1980s showed multiple times that the importance of a quantitative attribute increases significantly if it is defined on four levels versus two. But, does the number of levels effect continue to increase if we define that same attribute (holding range constant) on 11 or 21 levels? A secondary goal was to measure the strength of the number of levels effect for CBC studies involving modern design principles (e.g., modest level overlap).

Recall that we randomly assigned respondents to five different design cells (five different CBC questionnaires). The only thing different about the CBC questionnaires was how many levels we specified for the Price and Range attribute:

Cell 1: 2 levels
Cell 2: 3 levels
Cell 3: 4 levels
Cell 4: 11 levels
Cell 5: 21 levels

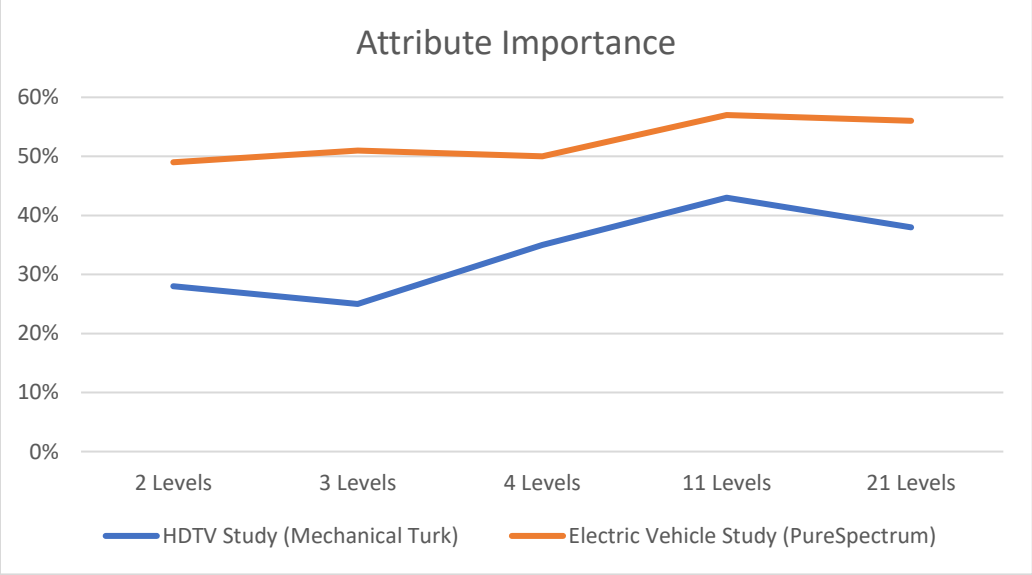
For each of our five CBC design cells, we estimated the Price or Range attribute as a single term using a linear function (constrained negative). We employed three different approaches for quantifying the importance of attributes: 1) importances summing to 100% from aggregate MNL, 2) importances summing to 100% from HB MNL, 3) importances summing to 100% derived from sensitivity analysis using a "what-if" market simulator built from the HB MNL utilities. For calculating the importance of an attribute (for methods 1 and 2 above) we used the standard approach of taking the utility difference between the best and worst levels. Detailed results for the three approaches are reported in the Appendix. Averaging across the three methods for estimating the effect of Price (for HDTVs) or Range (for EVs) on choices, we obtained the following Importance scores (Exhibit 3):

Exhibit 3

Number of Levels	Importance MTurk (Price)	Importance PureSpectrum (Range)
2 Levels	28%	49%
3 Levels	25%	51%
4 Levels	35%	50%
11 Levels	43%	57%
21 Levels	38%	56%

In an omnibus ANOVA followed by t-tests, the only significant difference we found in the MTurk column above is between the 3-level treatment of price and the 11-level treatment ($p<0.03$). Half of the differences in the PureSpectrum column are significant at the 95% confidence level (but note we had the advantage of larger sample sizes for this second study). We've plotted the same data below in a chart (Exhibit 4) for visual inspection:

Exhibit 4



(Note that sample sizes were larger for the electric vehicle study, so the importances have greater precision than for the HDTV study.)

Our MTurk findings (for HDTVs) seem to echo the earlier findings from academics and from Hoogerbrugge. Increasing the number of levels from two to four for a quantitative attribute seems to lead to increased importance in terms of effect on choice in conjoint questionnaires but does not continue to increase much beyond the use of four levels. However, our results for the electric vehicle study fielded using PureSpectrum sample suggest that the number of levels effect is modestly seen only when more than four levels are included in an attribute, but not for the range of variation between two and four levels. The number of levels effect seems, frustratingly, to vary from project to project. Perhaps the number of levels effect also depends on the quality of the panel source. Importantly, in neither of

our projects did the number of level effect come close to approaching the magnitude that Wittink and others reported decades ago for ratings-based full-profile conjoint. And, for neither project did the importance of the test attribute increase when moving from 11 to 21 levels.

Research into the number of levels effect suggests it is probably mostly or entirely due to psychological factors. To test the psychological hypothesis, we asked respondents both pre and post taking the CBC section of the survey to select which of the three attributes was most important to their decision in purchasing HDTVs and EVs. (We even encouraged respondents prior to the answering this question the second time to change their mind if after taking the CBC questionnaire their attitudes had shifted.) We saw very little difference pre and post on stated importance for Price and Range. Consistent with the psychological hypothesis for the number of levels effect, there was a slight tendency for respondents seeing Price or Range defined on two attributes to state that Price or Range was less important than other attributes (Post vs. Pre CBC measurement) compared to respondents who saw price defined on more than two levels. But the effect was not statistically significant.

Wrapping It Up

As much as we were hoping to shine a clear light on what to expect for CBC projects with the number of levels effect, our findings were not as clean as we had hoped. After conducting the HDTV study with MTurk respondents, we were encouraged and thought the story was going to fall in line with Hoogerbrugge's conclusions from 2000. But, the electric vehicle study fielded with higher quality PureSpectrum Panel surprised us, with a different and more favorably muted pattern manifested by the number of levels effect. What is comforting regarding both of our studies is that manipulating the number of levels from two clear up to 21 levels for a quantitative attribute led to relatively modest changes in the importance or impact of that attribute on product choice for CBC experiments. The number of level effect we find in 2023 for CBC studies is not nearly as strong as what Wittink and co-authors found in the 1980s and 1990s with ratings-based full-profile conjoint analysis. So, we conclude with the recommendation to specify quantitative attributes on three or more levels, which is likely to be more realistic in describing real-world market conditions better than just two levels of a quantitative attribute—and, to not fret so much about the number of level effect for CBC questionnaires.

Limitations of our research are that we examined just two CBC studies, and both studies employed three concepts per task designed with modest level overlap, plus a standard None concept.

References:

Currim, I. S., C. B. Weinberg, and D. R. Wittink. 1981. The design of subscription programs for a performing arts series. Journal of Consumer Research 8:67–75.

Hoogerbrugge, Marco. 2000. Practical Issues Concerning the Number-of-Levels Effect. In *Sawtooth Software Conference Proceedings*, pp. 113-123. Sequim, WA: Sawtooth Software.

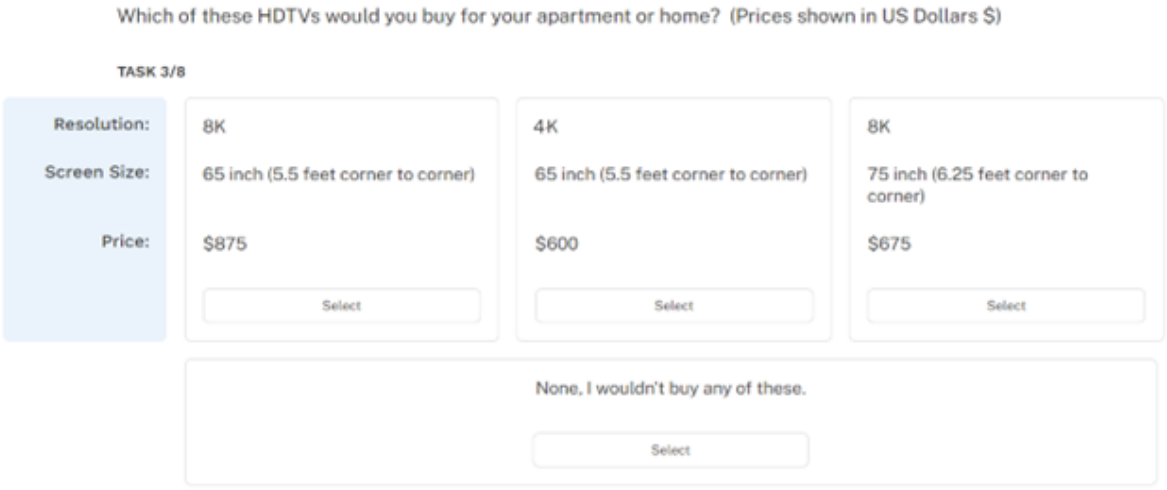
Orme, Bryan and Keith Chrzan (2021), "Becoming an Expert in Conjoint Analysis," Sawtooth Software, Provo, UT.

Orme, Bryan and Keith Chrzan (2022), "Real-Time Detection of Random Respondents in MaxDiff" accessed at: https://sawtoothsoftware.com/resources/technical-papers/categories/maxdiff-scaling

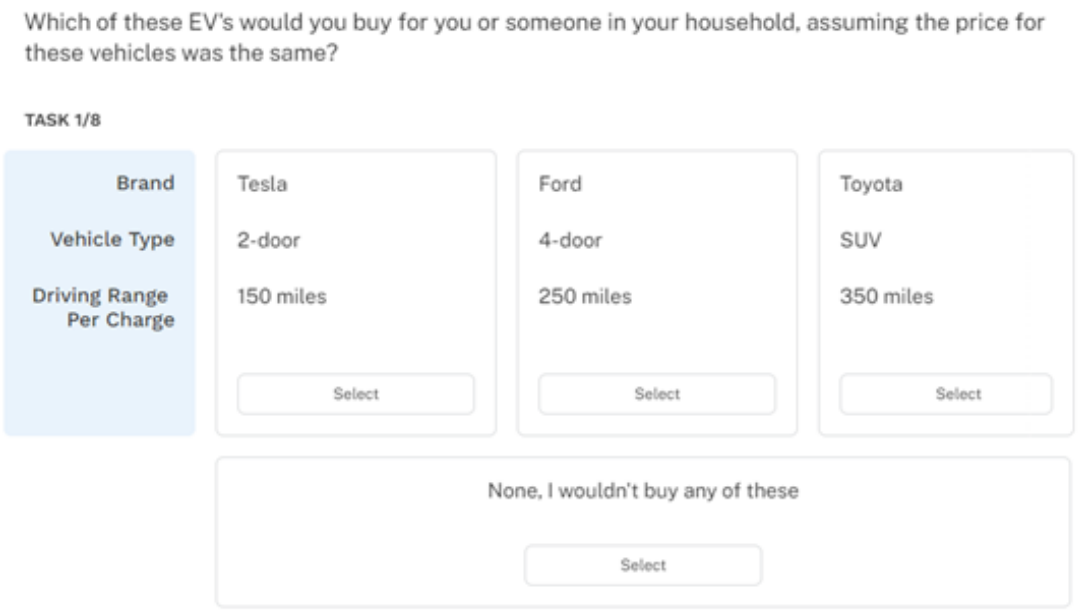
Wittink, Dick, Joel Huber, John Fiedler, and Richard Miller (1991), "The Magnitude of and an Explanation/Solution for the Number of Levels Effect in Conjoint Analysis." Working paper, as cited by Wittink in the 1992 Sawtooth Software Conference Proceedings.

Appendix A (CBC Choice Task)

HDTV Choice Task Sample



EV Choice Task Sample



Appendix B (Importances by Different Methods)

HDTV Importance Scores

Summary of Importances:

	CBC1	CBC2	CBC3	CBC4	CBC5
	Price 2 levels	Price 3 levels	Price 4 levels	Price 11 levels	Price 21 levels
Aggregate Logit:					
Resolution:	36%	32%	29%	26%	27%
Screen:	28%	43%	35%	27%	31%
Price:	27%	25%	37%	48%	42%
HB Individual-Level Importances:					
Resolution:	36%	35%	33%	28%	30%
Screen:	42%	47%	41%	39%	40%
Price:	23%	18%	27%	33%	31%
***Sensitivity Analysis on HB Utilities:**					
Resolution:	34%	32%	27%	27%	28%
Screen:	32%	37%	33%	24%	30%
Price:	34%	31%	40%	49%	42%
	n=110	n=108	n=116	n=115	n=126

EV Importance Scores

Summary of Importances:

	CBC1	CBC2	CBC3	CBC4	CBC5
	Range 2 levels	Range 3 levels	Range 4 levels	Range 11 levels	Range 21 levels
Aggregate Logit:					
Brand:	7%	12%	10%	8%	11%
Vehicle Type:	37%	28%	30%	25%	25%
Driving Range per Charge:	56%	60%	60%	67%	64%
HB Individual-Level Importances:					
Brand:	32%	30%	28%	25%	26%
Vehicle Type:	35%	33%	33%	33%	29%
Driving Range per Charge:	33%	37%	39%	42%	45%
***Sensitivity Analysis on HB Utilities:**					
Brand:	16%	12%	24%	16%	18%
Vehicle Type:	27%	31%	25%	21%	24%
Driving Range per Charge:	57%	57%	51%	63%	57%
	n=171	n=162	n=162	n=148	n=158

Note, for sensitivity analysis, we started with a base case product specification vs. a set of competitors and the None. One-by-one we changed the attributes across their levels and recorded the share of preference due to each feature change (holding all other attributes constant). We took the logs of the shares of preference associated with each attribute level change, treating them as if they were part-worth utilities, then computed importances in the standard way. The standard way to compute attribute importances is to take the range in utility from best to worst levels within each attribute, then scale these differences to sum to 100% across attributes. For HB utilities this is done at the individual level, then the importances are averaged across respondents.