Reprint citation: C Chapman and M Callegaro (2022). Kano Analysis: A Critical Survey Science Review. In B. Orme (ed.), Proceedings of the 2022 Sawtooth Software Conference, May 2022, Orlando, FL.

KANO ANALYSIS: A CRITICAL SURVEY SCIENCE REVIEW

CHRIS CHAPMAN
MARIO CALLEGARO
GOOGLE

ABSTRACT

The Kano method gives a “compelling” answer to questions about features, but it is impossible to know whether it is a *correct* answer. To put it differently, it will tell a story— quite possibly an *incorrect* story. This is because the standard Kano questions are low quality survey items, often paired with questionable theory and scoring. The concepts are based on durable consumer goods and may be inapplicable for technology products.

We follow our theoretical assessment of the Kano method with empirical studies to examine the response scale, reliability, validity, and sample size requirements. We find that Kano validity is suspect on several counts, and a common scoring model is inappropriate because the items are multidimensional. Beyond the questions about validity, we find that category assignment may be unreliable with small samples (N < 200). Finally, we suggest alternatives that obtain similarly compelling answers using higher quality survey methods and analytic practices.

INTRODUCTION

The Kano method (Kano, et al., 1984; Zacarias, 2015) is a relatively popular method to sort features into buckets related to their strategic appeal, such as whether a feature is unexciting but a must-have feature (“table stakes”) or it is something that users will like but don’t expect (and thus a “delighter”). In our experience there is one nice thing about Kano, but two areas of serious concern. The nice thing is that it puts features on a 2-dimensional landscape that is excellent for framing discussions with product managers, executives, and the like.

The two serious concerns are: (1) the usual Kano survey items violate several principles of survey design, and (2) the theory behind the Kano scores is dubious and has little empirical support. The resulting “story” will always appear plausible, but it is impossible to assess the results for validity. In short, we may be unable to trust whether the story is *true*.

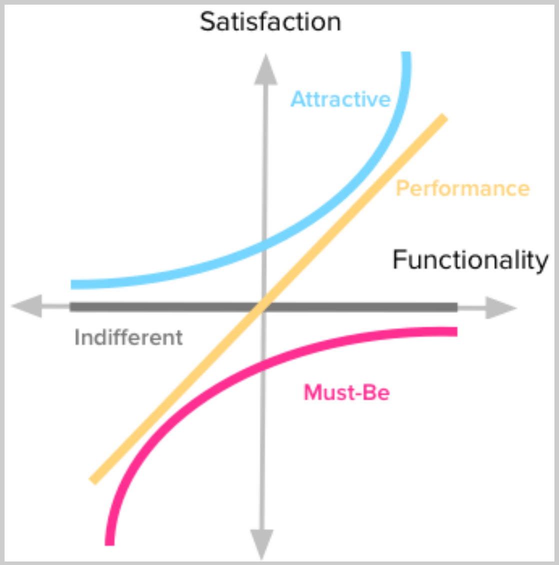
In this paper, we outline each of those concerns, review empirical results investigating them more deeply, and propose alternatives that are able to deliver the one “nice thing” while using acceptable and sound survey and analytic practices.

ONE NICE THING: TWO DIMENSIONS

Inspired by the two-factor theory of Herzberg et al. (1959) describing workplace motivation, Kano (1984) proposed that customer satisfaction with a product should not be conceived as a unidimensional construct, but rather as a two-dimensional construct. In our opinion, this was the primary insight that Kano emphasized, and it continues to be an appealing aspect of the model.

Given the assignments, features or products are placed on a **strategic plot on 2 dimensions**. This works well for stakeholder discussions! Figure 1 shows the typical Kano method dimensions and quadrants.

Figure 1: Two Kano Dimensions



(Image from Zacarias, 2015)

Concern 1: The Items, Wording and Scale

The typical Kano survey asks two questions about each feature under consideration:

1. How would you feel if your *[device, service]* worked this way or had this feature?
2. How would you feel if your *[device, service]* didn't work this way or didn't have this feature?

Respondents reply to answer each question with one of these categorical responses:

Rating scale (select one)

❍ I like it

❍ I expect it

❍ I feel neutral

❍ I can tolerate it

❍ I dislike it

Based on the responses to the 2 items, a product, feature, or service is assigned to one of 6 categories: "must have," "performance," "attractive," "indifferent," "reverse" (don't want), or "inconsistent." There are several schemes to map responses to categories (cf. Zacarias 2015). Following are canonical examples of mapping the 2 items to a category:

Category:	Example Item 1	+ Item 2:
Must have	Expect it	+ dislike not having it
Performance	Like it	+ dislike not having it
Attractive	Like it	+ expect not having it
Indifferent	Don’t care	+ don’t care about not having it
Reverse	Dislike it	+ like not having it
Inconsistent	Like it	+ like not having it

From a survey design perspective, there are several problems with these items:

- The questions concern **hypothesized** feelings in the *future*, with regards to a **hypothetical situation**. Such questions tend to have low reliability and are almost impossible to assess for validity. (What would count as evidence of consistency for a future hypothetical?) Better is to ask about *current* attitudes or behavior.
- The second question asks about a **negative/absence** situation. This is very difficult for a user to answer— *except* for products they already use (which is not a common situation for such surveys by UXRs). Also, negative direction items tend to have lower reliability.
- The **rating scale** has items that are *non-mutually exclusive* yet are presented as exclusive categories. Using mutually exclusive response options is a standard suggestion in the question wording literature, and not doing so is considered a common mistake (e.g., Bolton & Brace, 2022, p. 114; Krosnick & Presser, 2010, p. 264; Lavrakas, 2008). This also leads to lower reliability and potential respondent inconsistency. Better is to rate each attitude separately rather than as a single forced choice.
- The scale conflates **multiple dimensions**, yet is presented as a single, nominal scale.

In short, the response scale does not make sense. We can see this clearly in a specific example of the scale problem. Instead of exclusive responses, it may be perfectly reasonable to select *more than one*, or potentially even *all* responses. Let’s imagine that the question is the following:

How would you feel if Apple iPad had a higher-performance LightningX cable?

One could quite rationally answer “yes” to every response, and might select any answer depending on mood, the survey item order, or available time to take the survey:

☒ I like it — *yes, because it is faster*
☒ I expect it — *yes, because Apple makes such changes all the time*
☒ I feel neutral — *yes, because I can see pros and cons of it*
☒ I can tolerate it — *yes, because I can buy new cables*
☒ I dislike it — *yes, because overall I’d prefer* not *to switch*

The items are not mutually exclusive, and therefore Kano scoring is not a reasonable way to assess the responses (that’s after *assuming* that the questions are reliably answered). The items inappropriately map multiple dimensions to a single nominal response (which is later treated in scoring as an *ordinal* response). In the empirical results below, we demonstrate that the response scale is multidimensional and is poorly approximated by a unidimensional model. In other words, it is a bad idea to have mutually exclusive options on this scale.

You might wonder, “Isn’t it OK to force users to choose which feeling they have most?” First, why do that? Why not assess all their feelings? Second, that is not actually the question it is asking, and respondents may be confused by the scale. In the empirical results, we find that respondents often give contradictory and unreliable responses.

More importantly, what theoretical grounds would we have to force a tradeoff among these options? If it is a tradeoff, does Kano model the data as a tradeoff? Unfortunately, Kano theory does not have an answer; and the analytics do *not* model the data as tradeoffs. Kano simply assumes that it is OK to require a single forced choice among multidimensional items.

Recommendation for response scale: Instead of the Kano questions and response scale, use items that are about *current behavior*, that separate *multiple dimensions*, and that use *non-exclusive response* options. (*Note*: this would require changing many Kano scoring models.)

Concern 2: Theory and Scores

There is no known validity of the four quadrant concepts of “delighters,” etc. There is a large literature of papers that use Kano yet almost all of them are case studies of single applications, with little or no discussion of whether the method is *reliable*, *valid,* or *replicable*.

Brief Literature Review

In the original paper (Kano et al., 1984) there is no claim that the questions or scoring method are correct or generalizable. Instead, it argues that the attractiveness of a product should be considered in 2 dimensions instead of 1 dimension and discusses 2 examples.

Four review papers have examined the larger literature related to Kano. One review paper (Löfgren and Witell, 2008) notes the following: “A review of 33 papers relevant to the theory of attractive quality revealed several developments with respect to methodological issues, but *many of these lack the scientific basis that would justify inclusion* in the theory.” [italics added]

A second review (Hartmann and Lebherz, 2016) noted, “*research content* [about the Kano model] *has to focus more on the theory* and its implications itself. Instead of doing that, many contributions are recently applying the Kano model in specific contexts without questioning the implications. Other examples are modifying the model, without showing the differences and implications in detail” [italics added]. In other words, there is not a systematic method but multiple approaches with conflicting theories and methods.

A third review (Mikulić, 2007) summarized 46 papers applying the Kano model and variations. It noted that the model was “well adopted,” although despite the popularity, “Nevertheless, at present, there is still no clear consensus among researchers about the most appropriate assessment method, and convergent validity between the different methods has not been confirmed yet” (p. 7). It finds that authors of the various research papers disagreed to some extent about nearly all aspects, including the underlying theory, dimensions, items, scale, and scoring procedures. This suggests that the label “Kano” may be better regarded as an inspirational *family* of practices rather than a method as such.

A final review paper, (Witell, Löfgren, and Dahlgaard, 2013), considered 147 research papers. They found an "explosion" of applications, but that "too much research has simply applied the Kano methodology without discussing its implications for the theory" (p. 13). They concluded that "it is now necessary to revisit the theoretical foundation of the theory of attractive quality . . . little has been done in terms of enhancing our knowledge of the theoretical similarities and differences of the concepts of satisfaction and dissatisfaction" (pp. 17–18). In short, the theory remains undeveloped and is uncertain even with regard to foundational elements such as the axis of satisfaction.

Conclusion for Theory

The original Kano paper and the few reviews that examine the method agree with the one "nice thing" we note above about multidimensional assessment. They do not demonstrate strong support, or even general agreement, for the basic *theory* of the model.

Now, one might wonder, "*We used it for some project, and it worked well. Isn't that evidence?*" How do we know that it worked well? Were the answers compared to another method? Or was it simply assumed to work because stakeholders liked the answer? This is an example of asking something we would like to know, but customers cannot truly answer (cf. Chapman 2013).

Empirical Studies: Reliability and Validity

Data

Data were collected from N=10,638 respondents using Google Surveys, in a total of 7 survey versions (see the Appendix for example screenshot). Respondents answered Kano items for the following 3 features (which do *not* represent Google product research, analytics, or feature plans; they were selected by the authors for salience):

- Imagine your next phone has a **touch screen**
- Imagine your next phone **recharges with no cable**, using environmental light and motion
- Imagine your next phone has a higher megapixel, **higher resolution camera**

The intention was to have one feature that was expected to be categorized as a "must have" (touch screen); another that was a "performance" feature (higher resolution); and one that would be an "attractor" (cordless recharging). In one version (N=1501) we additionally asked one of the items *twice*, in order to examine within-subject item reliability.

In this report, we use those data to investigate several questions:

- Are the Kano items reliable? Are they valid and consistent with Kano theory?
- Is the Kano response scale unidimensional, as it is typically used?
- Do the results align with expectations about the features?
- Are the results stable? Are they expected to be stable in qualitative settings?

Question 1: *Are the Items Reliable?*

A: No. The levels of raw agreement—as well as the reliability coefficient for a repeated item—were lower than accepted standards for "good" items.

Details. First, we examine the degree to which respondents will simply give the same answer twice, when asked a few seconds apart within a single study. We observe the following, where the rows present agreement proportions between the first time (row) and second time (column) that the item was asked. (The most common response when asked again is shown in **bold**).

	I can tolerate it	I dislike it	I expect it	I feel neutral	I like it
I can tolerate it	**0.5441**	0.0294	0.0294	0.2059	0.1912
I dislike it	0.0500	**0.7750**	0.0250	0.0500	0.1000
I expect it	0.0541	0.0135	0.3919	0.1351	**0.4054**
I feel neutral	0.0438	0.0255	0.0292	**0.7482**	0.1533
I like it	0.0134	0.0077	0.0239	0.0450	**0.9100**

For example, for respondents who replied "I can tolerate it" when first asked, they later replied "I can tolerate it" at a proportion of 0.5441 (54% of the time). 46% of the time, they gave different responses. The worst level of agreement was for "expect"—only 39% of respondents gave the same answer a few seconds later, whereas 40% changed their answers to "like."

What does this mean for the Kano scoring? Responses of "expect" and "like" are scored quite differently in the Kano model—a response of "like" on the first item most commonly aligns with attractors, while "expect" never does (scoring in the Folding Burrito scheme). Because responses are highly inconsistent between those two choices, we conclude that (at least in these data) ***Kano score assignments would be expected to be erratic and inconsistent.***

For example, in our data, the "touch screen" feature received N=5240 responses of "Expect" (68.7% of the responses). We did not test the reliability of "expect" for that feature—but, based on the feature where we did test it ("no cable"), we would expect 61% of the respondents could have given a different answer, mostly "like." In that case, instead of being a "must have" feature (see below), it likely would not have been scored as "must have." (For example, if 40% of those responses switched to "like," then "like" would dominate—and "must have" is not a possible category outcome in that case.)

The key point is this: there is strong evidence that Kano responses are unreliable; yet the scoring model that assigns categories does not take that into account. Given the finding above (and next), we suggest that Kano item reliability should be explicitly tested and assessed for every feature as a precondition for assuming that the items can be scored appropriately.

Second, we can calculate the actual degree of agreement between the answers. There are several ways to assess this. A common statistic for direct comparison is the adjusted Rand Index, which takes into account the base rate of the answers (e.g., the fact that so many responses for all features are "like"; Chapman & Feit, 2019, p. 330). For these data, that is ARI = 0.611. This says that the responses are 61% "better than random agreement" (or, conversely, 39% worse than perfect agreement, after a few seconds' time). Put differently, once we remove the acquiescence bias (the high base rate to respond with "like"), respondents agreed with themselves—in responses given a few seconds earlier—only 61% of the time.

If we regard the data as ordinal or continuous (as suggested, e.g., in the Folding Burritos guide), then we may compute a correlation coefficient. The Kendall *tau* correlation coefficient applies to non-parametric, ranked (ordinal) responses, while Pearson's *r* may be used for continuous data. In this context, that is often referred to as a reliability coefficient. For these data, we observe Kendall's *tau* of 0.70 (note that *tau* does not have confidence intervals):

We find Pearson's *r* (treating responses as continuous; coding as noted in the next section) of r = 0.735 (95% confidence interval of 0.711–0.757). A typical target for a "high" reliability item would be r > 0.90. A "good" item might have r = 0.80–0.90. A "marginal" item might have r = 0.70–0.80, while a "poor" item would have r < 0.70. In these data, the Kano item is in the lower range of "marginal," according to that standard.

Conclusion for item reliability: There is a high rate of response disagreement when a Kano item is re-asked after a short delay. In these data, the items did not demonstrate good reliability. Answers disagreed sufficiently to pose questions about stability of the Kano category scoring (see the next few sections).

Question 2: *Do the Items Associate as Expected, if Kano Theory is Valid?*

A: No. The items had unexpected (and universally negative) patterns of correlation with one another, which is generally contrary to Kano theory.

Details. Kano scoring assumes that the scaled items are on an ordinal or quasi-continuous scale. For example, the author of the Folding Burritos guide (https://foldingburritos.com/kano-model/) proposes the following values for the scale responses:

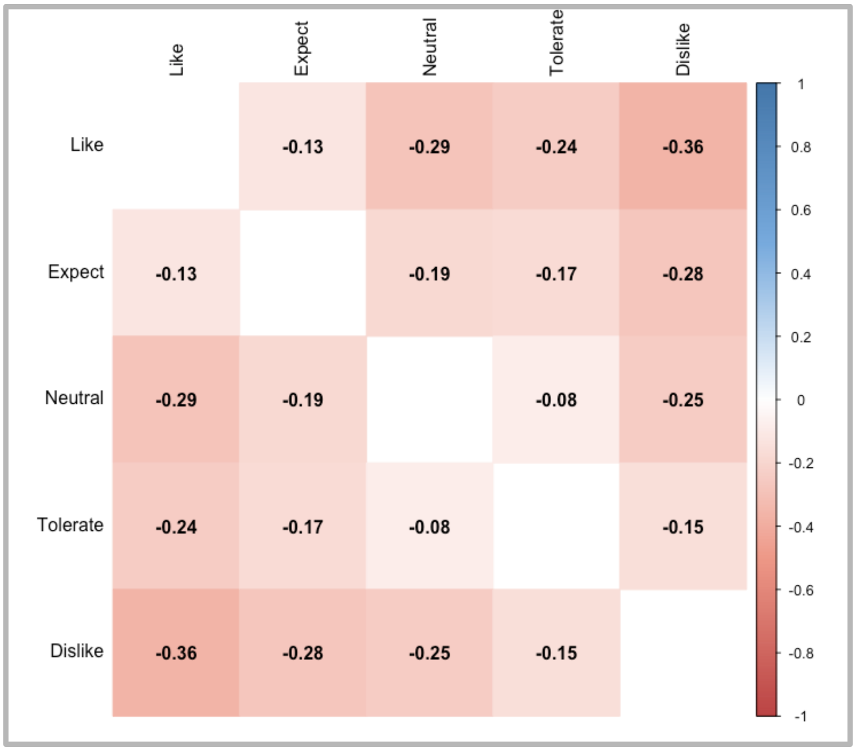
Functional: -2 (Dislike), -1 (Live with), 0 (Neutral), 2 (Must-be), 4 (Like);
Dysfunctional: -2 (Like), -1 (Must be), 0 (Neutral), 2 (Live with), 4 (Dislike);

If so, we would expect to see particular patterns in Kano data. To test this, we ran 2 versions of the survey in which multiple responses were possible. For example, a user might check both "like" and "expect"—and thus, we can evaluate the correlation between those responses. Following are the *expected* patterns by Kano theory, and an interpretation of the *results* (see Figure 2 for the correlation coefficients):

	Expectation	**Result**
1.	Positive associations:	
	a. r(Like, Expect) > 0	No
	b. r(Dislike, Tolerate) > 0	No
2.	Negative associations:	
	a. r(Like, Dislike) < 0	Yes
3.	Ranked correlations as follows:	
	a. r(Like, Expect) > r(Like, Neutral)	Yes
	b. r(Like, Expect) > r(Like, Tolerate)	Yes
	c. r(Like, Expect) > r(Like, Dislike)	Yes
	d. r(Like, Neutral) > r(Like, Tolerate)	No
	e. r(Like, Neutral) > r(Like, Dislike)	Yes
	f. r(Like, Tolerate) > r(Like, Dislike)	Yes
	g. r(Expect, Neutral) > r(Expect, Tolerate)	No
	h. r(Expect, Neutral) > r(Expect, Dislike)	Yes
	i. r(Expect, Tolerate) > r(Expect, Dislike)	Yes
	j. r(Neutral, Tolerate) > r(Neutral, Dislike)	Yes

In short, of 13 associations we would expect to find, we confirmed 9 associations in the same direction as Kano theory. However, note that of those 9 associations, *all except 2 cases of agreement with theory involved the "Dislike" item*. Acquiescence bias may account for respondents disfavoring the negative end point of a scale.

Figure 2: Correlation Matrix Plot for Item Correlations in the Kano Survey



The numbers shown are Pearson's r correlation coefficients for pairs of items. The overall pattern is perhaps most consistent with acquiescence bias (such as a dispreference of respondents to answer "dislike"). Overall, the individual comparisons—such as the negative correlation of "like" with "expect"—do not align well with assumptions of the Kano model.

What do we see if we remove the tests that involve "dislike"? Here are those results:

1. Positive associations:
   a. r(Like, Expect) > 0 — No
2. Negative associations:
   a. n/a
3. Ranked correlations as follows:
   a. r(Like, Expect) > r(Like, Neutral) — Yes
   b. r(Like, Expect) > r(Like, Tolerate) — Yes
   c. r(Like, Neutral) > r(Like, Tolerate) — No
   d. r(Expect, Neutral) > r(Expect, Tolerate) — No

After removing the unique character of the Dislike item (see the factor analysis section below), we find that only 2 of the 5 tests align with Kano theory. One of those results is that, quite surprisingly, the Like and Expect items have negative correlation.

Conclusion for Kano item intra-survey validity assessment. The pattern of item associations is largely inconsistent with Kano theory and assumptions.

QUESTION 3: *IS THE KANO SCALE UNIDIMENSIONAL, AS IS ASSUMED? IS IT OK TO GET A SINGLE RESPONSE ON THE SCALE?*

A: No. The Kano scale appears to conflate at least 2 dimensions (possibly 4 or 5). We recommend using a multi-dimensional response option instead (see the “alternatives” above). We would not use the Kano response scale, nor would we use a scoring model that assumes a unidimensional scale (such as the Folding Burritos values assigned to the scale points: https://foldingburritos.com/kano-model/).

Details. Using the multiple response data (see previous question), we performed exploratory factor analysis (EFA) on the correlation matrix. First, we used the R nFactors library to examine the most likely number of factors, according to several methods; this suggests that the most likely number would be 1, 2, or 4 factors.

We then used the R factor analysis procedure to examine the 1-factor and 2-factor solutions. (*Note*: it is not possible to use that procedure for 4 factors, because there are only 5 items—that is essentially the same as saying that every item is a separate factor, plus a degree of freedom for error variance—thus, there are no “factors,” just items.) In the case of 2 factors, we allow the factors to be non-orthogonal, using “oblimin” rotation. Those results are:

1-Factor Solution

```
Loadings:
              Factor1
Like          0.365
Expect        0.278
Neutral       0.245
Tolerate      0.152
Dislike       -0.998
              Factor1
SS loadings   1.288
Proportion Var 0.258
```

The extracted factor is nearly identical (loading=0.998) with the *Dislike* item. There is also a moderate and positive correlation for the *Like* item, suggesting that *Like* and *Dislike* are *not* likely to be modeled as opposites as asked in this item. (Caveat: because respondents could endorse any number of responses, there may be an induced correlation due to response style. Future research might explore this, using multiple, single dimensional items.) Apart from *Dislike*, none of the other items loads highly on the factor—it is just a “dislike factor.” Overall, the model captures 25.8% of the modeled (non-error) variance.

We evaluate the one-factor solution as relatively poor and uninteresting. However, a one-factor model is consistent with the hypothesis above that *Dislike* responses may be driven by a unique process (such as acquiescence). This differs from Kano assumptions. If one wishes to assess this factor, we recommend instead to use a scale optimized to assess *disliking*.

2-Factor Solution

Loadings:

	Factor1	Factor2
Like	**-0.982**	
Expect	0.355	0.242
Neutral	0.375	0.407
Tolerate	0.255	0.325
Dislike	**-0.982**	

	Factor1	Factor2
SS loadings	1.299	1.296
Proportion Var	0.260	0.259
Cumulative Var	0.260	**0.519**

Factor Correlations:

	Factor1	Factor2
Factor1	1.00	**0.27**
Factor2	0.27	1.00

This 2-factor solution shows stronger fit to the data, capturing 51.9% of the modeled variance. However, it is still not a very interesting factor model. It says, in effect, that there are 2 separate dimensions—*Like* and *Dislike*—with the other items not aligning well to either dimension (suggesting that there might be additional dimensions that are still not captured). Also, the factors of *Like* and *Dislike* have relatively low (r=0.27) correlation; they do not form anything like opposites on a single dimensional scale.

If we take just the "Like factor" (factor 2), it is worth noting that the other Kano items do not load on it as expected; the correlations are contrary to Kano theory. For example, we would expect loadings with *Like* to be as follows: *Expect >Neutral >Tolerate*, *Expect >>Tolerate*, and the loadings for *Expect* should be positive while *Tolerate* should be negative.

However, the *Expect* item has a negative loading (reversing signs in the loading table, because of the negative loading of *Like*, which is an algebraic quirk of no importance). *Tolerate* loads more strongly than *Expect* (loading=0.325 vs. 0.242), while *Neutral* is higher than either of them (0.407). Of the 5 implied tests of consistency with Kano theory (*Expect >Neutral*, etc.), the data contradict 3 of the 5 tests (*Expect*!>*Neutral*; *Expect*!> *Tolerate*; *Expect*! = positive).

Possible future work: Use Confirmatory Factor Analysis to test the 1-factor and 2-factor models for goodness of fit on separate data sets. (This would require new data for replication.) Expected outcome: there is no reason to expect that the 1-factor solution would be preferable.

Conclusion for factor analysis: In these data, the Kano scale is *not* unidimensional. It is better modeled as 2 or more separate factors. When that is done, the factor structure is not consistent with Kano theory. We do not recommend the common Kano scale or continuous scoring of it.

QUESTION 4: *DO THE RESULTS ALIGN WITH EXPECTATIONS ABOUT THE FEATURES' KANO CATEGORIES?*

A: Partially. 2 out of 3 features were categorized as expected. With N=7624 responses, the following features for a "new smart phone" were categorized:

Feature	**Expected Category**	**Result in Data**
Touch screen	Must Have	Must Have (59.9%)
Higher resolution camera	Performance	[none; largest=Attractive, 30.5%]
Cable free charging	Attractive	Attractive (56.8%)

Following are breakdowns of categories for the 3 features tested.

Touch Screen

Attractive	Indifferent	**MustHave**	Performance	Questionable	ReverseInterest
0.054	0.111	**0.599**	0.194	0.036	0.006

Higher Resolution

Attractive	Indifferent	MustHave	Performance	Questionable	ReverseInterest
0.305	0.244	0.159	0.227	0.057	0.008

Cable-Free Charging

Attractive	Indifferent	MustHave	Performance	Questionable	ReverseInterest
0.568	0.219	0.022	0.085	0.073	0.034

Of course, this might be regarded as learning something—we might have been wrong in our expectations about the features' categories and learned something new about customers' perceptions. That would be a more convincing argument if the items had internal consistency with Kano theory (see the theoretical discussion above).

Beyond that, although the results aligned with expectation for 2 of the 3 features, there were additional concerns in the data. First, there are a modestly high rate of unexplainable responses, due to both inconsistency ("questionable" responses in 4–7% of data) and because they do not align to expectations (e.g., 19% responses that a *touch screen*—presumably a requirement for any smartphone—is a "*performance*" feature, with another 11% indifferent).

Second, we see that even in this simple case, it is difficult to interpret why respondents gave a particular response. Consider "higher resolution camera," which we might assume is very nearly a classic example of a *performance* feature. Only 22.7% of respondents classified it as performance (based on the Kano scoring). But we see that 24.4% were *indifferent* and 15.9% were classified as regarding it as a *must have*. It makes little sense for "higher resolution" to be a *must have* feature, unless they do not currently have a smartphone.

Discussion of Category Assignments

The inconsistency in category assignment poses a serious question for the application of Kano concepts with technology products. For technology products, continually increasing performance is generally assumed by users, and users often purchase products to get higher performance (battery life, speed, etc.). But in that case, is there a meaningful difference between a “performance” feature, a “must have” feature, or an “attractive” feature? Customers might rightly say, “of course performance must increase” (and thus performance==must have), or “I am attracted to higher performance” (and thus performance==attractor==must have).

This suggests that the core theoretical constructs of *performance*, *attraction*, and *must have* may not be distinguishable or make sense for technology products. We believe that the underlying Kano theory—and how it maps to the questions and scale—is extremely unclear. Empirically, our results align with that concern (e.g., the unclear category assignments for a higher performance smartphone camera).

Conclusion for category assignments: 2 of 3 features aligned with expectation. However, the data suggest additional questions about the validity of Kano theory for rapidly changing technology products. Technology products may not align well with the assumed Kano concepts of performance, attractive, and must-have features.

QUESTION 5: *ACROSS THE LARGE SAMPLES, IS THERE AGREEMENT ON THE KANO CATEGORY ASSIGNMENTS FOR THE FEATURES?*

A: Yes, in samples ranging N=1501–1611, the results were stable, testing the “resolution” feature.

Details. Following are replications of category assignment for the “resolution” feature (see previous question), in 5 samples. In each one, we highlight the top two most commonly assigned Kano categories.

Resolution Sample 1

Attractive	**Indifferent**	MustHave	Performance	Questionable	ReverseInterest
0.313	**0.260**	0.147	0.221	0.053	0.007

Resolution Sample 2

Attractive	Indifferent	MustHave	**Performance**	Questionable	ReverseInterest
0.329	0.227	0.133	**0.246**	0.058	0.006

Resolution Sample 3

Attractive	**Indifferent**	MustHave	Performance	Questionable	ReverseInterest
0.303	**0.292**	0.114	0.224	0.061	0.006

Resolution Sample 4

Attractive	Indifferent	MustHave	**Performance**	Questionable	ReverseInterest
0.301	0.210	0.208	**0.224**	0.046	0.011

Resolution Sample 5

Attractive	**Indifferent**	MustHave	Performance	Questionable	ReverseInterest
0.**279**	0.**230**	0.193	0.220	0.066	0.011

In all 5 samples, "attractive" was the most commonly assigned category, scoring respondents' answers. This suggests that for large samples (N ~ 1500), the Kano category assignments replicate well. On the other hand, as we noted above, the modal "winning" category assignment may not be a good representation of users. In this case, "attractive" was the modal category in all 5 samples, yet 67–72% of users assigned it to a different category in each sample!

Additionally, the assignments may be stable and yet not align with the presumed Kano theory. See above for more; briefly, we see here that a presumed canonical *performance* feature was never categorized as being a performance feature (as its first-place assignment). In only 2/5 samples was it categorized in its *second* most likely result as a performance feature.

Conclusion for large sample stability: Kano assignments are similar across repeated large samples. If you accept the theory and scoring (see above), then the results are expected to replicate, if based on a large sample (N=1500 here; but likely N=200, see below).

QUESTION 6: *ARE THE KANO CATEGORY ASSIGNMENTS STRONG ENOUGH TO USE FOR BUSINESS PURPOSES? (SETTING ASIDE QUESTIONS ABOUT VALIDITY)*

A: Unclear. In these data, the percentages of respondents giving the most common (modal) category response are rather low. Thus, the assigned, modal Kano category represents relatively few respondents—as few as 28% of the sample. When as many as 72% of users categorize a feature differently than the category that might be reported to business stakeholders, it poses concern about the appropriateness of acting on that modal category.

Details. Let's consider the "higher resolution" and "no cable" features. Let's look at the final sample for each feature. For "higher resolution" we see this distribution of answers:

Resolution Sample 1

Attractive	**Indifferent**	**MustHave**	**Performance**	Questionable	ReverseInterest
0.279	**0.230**	**0.193**	**0.220**	0.066	0.011

For "higher resolution," it is difficult to assert confidently that it is an "attractor" feature when 72.1% of respondents disagree!

For "no cable" we see the following distribution:

No Cable Sample 1

Attractive	**Indifferent**	MustHave	Performance	Questionable	ReverseInterest
0.572	**0.215**	0.023	0.075	0.075	0.041

"No cable" was rated as an attractor by 57.2% of respondents. That is a stronger finding than the "resolution" preference—but is it strong enough? Maybe. However, that is only one of these 2 features. Overall, we believe that neither is strong enough to conclude that a single Kano assignment—as might be made on a plot—is strong enough to be presented as "the category."

It is also important to examine which users aligned with which category. For example, although only 21% of users assigned "no cable" to the Indifferent category, it might be that these are the most important users for our product (such as potential early adopters, who may have a unique view or understanding of the feature).

What do we suggest? Consider an alternative method described above. If one uses the Kano method, then report the *distribution* of category assignments, i.e., the proportion of users aligning with each of the possible categories. Do not report only a modal category.

Conclusion for business usage of category assignment. Features are unlikely to align cleanly with a single, modal category. It appears likely that users will disagree substantially. Assuming—of course—that an analyst believes that the Kano theory and scoring are valid and reliable, we believe it is important not to report the outcome as a single modal category assignment for a particular product or feature. Instead, use a large sample and report the *degree* to which a feature aligns with every category, across the sample (see next section).

QUESTION 7: *IN SMALL-SCALE SAMPLES DRAWN FROM THESE RESPONDENTS—SUCH AS WE MIGHT OBTAIN IN QUALITATIVE RESEARCH STUDIES—DO THE RESULTS REPLICATE ACROSS SAMPLES?*

A: No. In these data, a sample size of N=200 is required for acceptable replication of Kano assignment. We believe this may be a *minimum bar* (especially when testing more than 3 features). Note that this sets aside the question of whether the assignment was *valid*; it only tests whether the assignment *replicated*, even if the implications are invalid.

Details. For the 3 features tested, we examined the consistency of the Kano assignment in 1000 repeated random samples for each feature, using different sample sizes. From the overall data set of N=7624 observations, we drew smaller samples with N=10 responses, and then increasing the sample size—with 1000 iterations each—to N=12, 15, 20, 30, 50, 100, 200, 300, and 500 responses.

We would propose that a "consistent" answer is one that is correct (compared to a different sample) for each feature at least 80% of the time.

Doing so, we find the following prevalence rates for the most frequent Kano-assigned category for each of the three features, by sample size. In this table, f1 = "touch screen" (*must have*), f2="higher resolution" (assumed to be a *performance* feature, but categorized most often as attractive), and f3="cable free charging" (**attractive**).

Sample Size	f1	f2	f3	average	combined
10	0.943	0.501	0.938	0.7940000	0.4431515
12	0.954	0.508	0.930	0.7973333	0.4507078
15	0.969	0.529	0.967	0.8216667	0.4956852
20	0.987	0.546	0.980	0.8376667	0.5281240
30	0.996	0.605	0.993	0.8646667	0.5983619
50	1.000	0.653	0.999	0.8840000	0.6523470
100	1.000	0.767	1.000	0.9223333	0.7670000
200	**1.000**	**0.854**	**1.000**	**0.9513333**	**0.8540000**
300	1.000	0.904	1.000	0.9680000	0.9040000
500	1.000	0.966	1.000	0.9886667	0.9660000

As an example, we may read the results for the N=12 row as follows. Feature 1 was assigned to its most common category in 95.4% of the 1000 samples, whereas Feature 2 was assigned to its dominant category in only 50.8% of the 1000 samples. The overall rate for the 3 features combined was a 79.7% average in correct assignment rate and combined (all 3 features) rate of 45.1% correct. This means based on these data, if we sampled 12 respondents, we would expect overall to be correct about any given feature (compared to a different sample) 79.7% of the time; and would be correct about 1 of the 3 features only 50.8% of the time, and all 3 features 45.1% of the time.

Thus, we expect that N=12 would not achieve an 80% level of accuracy for each feature (it would, instead, be expected to be 45% accurate for every feature, and as low as 51% for some feature—but we wouldn't know which one—compared to a new sample). We notice that even with N=50, the combined accuracy is only 65%. Thus, if we conducted 2 studies with N=50 in each one, we might expect different results 35% of the time for at least 1/3 features.

How many respondents do we need for 80% accuracy? In these data, we find that 80% accuracy is achieved at the level of N=200 or more respondents. At that sample size, F1 and F3 were 100% accurate (compared to new samples), while F2 was accurate 85% of the time. At N=200 we also find that the overall combined accuracy was 95% on average, and 85% for the set of all 3 features. In short, we might expect to be "right" >80% of the time when testing 3 features, if we have N >= 200 respondents—larger than a typical qualitative study.

Important Caveat. This finding is likely to be highly influenced by the exact items tested, and how many are tested. For example, if you test more than 3 items, you would be more likely to have one fall below the 80% rate (just because there are more chances). Also, if an item is unclear, it may be less likely to have consistent assignment. In the present data, we regard 3 features as a small number of features to test, and "higher resolution camera" as a relatively clear and understandable feature. Most Kano studies would have at least 3 and often more features; and the features are likely to be less clear than "higher resolution camera." Thus, we believe these results are likely to be a *lower bound*—i.e., minimum requirement—for the sample size needed for consistent results.

Conclusion for qualitative sample sizes. We believe samples of N=200 are required for consistent (if not necessarily valid) results using Kano questions and the Kano response scale. With N < 200, we would expect at least 1/3 of features to be miscategorized more than 20% of the time. With N <= 15 respondents, the overall accuracy rate may be less than 50%.

Summary of Empirical Results

The standard Kano items do not appear to meet standard criteria for reliable survey items, and response patterns cast doubt on the validity of the underlying theory. The standard response scale is not unidimensional, and it may be inappropriate to use it in that way. The assigned categories—setting aside whether they are valid—are not stable in small samples, so the method does not appear to be suitable for qualitative (small N) studies. A minimum N=200 respondents—or more, if testing many features—are likely to be needed for consistent results.

These may be minimum bars. These results were obtained in a relatively clear product space (smartphones), testing only 3 features. In less clear product spaces, or when testing more than 3 features, we would expect lower item validity and higher inconsistency of results. However, that is an empirical question. Thus, before using Kano in a new product space, we would suggest performing reliability and validity analysis to assess the quality of the results, along with the appropriateness of the target sample size.

Alternatives to the Kano Method

There are a variety of options that deliver a compelling 2-dimensional strategic map similar to the results of a Kano study, while using more reliable and standard survey procedures. An overall approach is this: assess the exact same list of items or products with 2 or more dimensions and plot them. That might be done with traditional scales and/or with MaxDiff tasks.

If you have 2 dimensions of interest—such as *importance* + *satisfaction*, or *importance* + *willingness to pay*, or *preference* + *brand affinity*—then we often recommend a MaxDiff for preference plus a Likert type rating scale for the other dimension. If you have *several* dimensions (such as brand personality dimensions), then I recommend a composite perceptual map that will relate them all to a convenient 2-dimensional plot.

Alternative 1: Plot 2 Likert Scales vs. One Another

If you want to align features on 2 dimensions, you can ask about those 2 dimensions, using 2 Likert (or other) scale survey items. For example, suppose that you want to assess the [imagined] LightningX cable for a phone, and you believe that there might be a mixture of attractiveness due to the *performance* and reluctance due to the *compatibility* of cables and devices that people have. You might write 2 Likert type items to assess those dimensions.

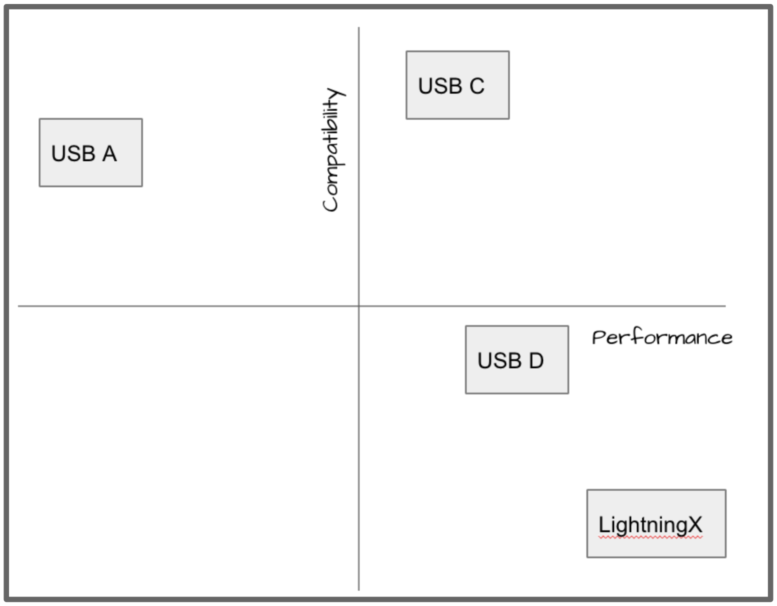
Following is an example of potential item wording. Note that this is just an example; it should be adapted and pre-tested for any specific problem.

> **Introduction**
> *(materials for the user that briefly review the LightningX concept)*
>
> **Q1. LightningX (LX) cables would have higher performance. Which of these best expresses your opinion about LX performance:**
>
> LX high performance does not appeal to me at all
> LX high performance slightly appeals to me
> LX high performance moderately appeals to me
> LX high performance extremely appeals to me
> I do not know enough to say
>
> **Q2. LightningX would not be compatible with existing cables. Which of these best expresses your opinion about LightningX compatibility:**
>
> LX cable compatibility is not a concern for me at all
> LX cable compatibility is a slight concern for me
> LX cable compatibility is a moderate concern for me
> LX cable compatibility is an extreme concern for me
> I do not know enough to say

If you have other concepts (such as USB A or C cables, or some other new concept LightningZ or USB D) then you could ask similarly about those. Then calculate (for example) the mean score on the 2 items, and plot those against one another for the various concepts. This might look like Figure 3 (*fake data*).

In Figure 3 (fake data), we see that USB A would be a non-starter, with worse performance and compatibility than USB C. USB D is probably not worth it from a user point of view— it has slightly higher perceived performance, but much worse compatibility. LightningX has strong performance appeal, but we would want to do as much as possible to help with users' concerns about compatibility. The two axes are flexible—you might use exactly the same as Kano, but also might adapt them to fit your product and business questions.

Figure 3: Hypothetical Scoring of Likert Style Responses on 2 Dimensions

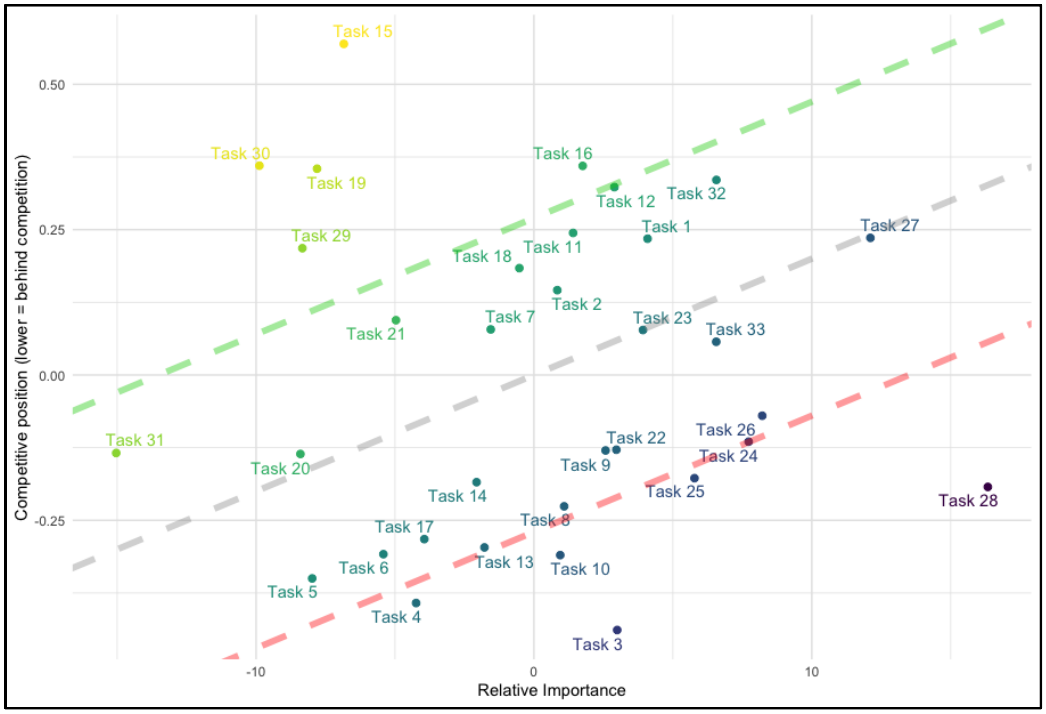


Alternative 2: MaxDiff Plus a Rating Scale or a Second MaxDiff

It is typical for one axis in a Kano-type chart to be the feature "importance." If you have many features, then MaxDiff is a good way to assess the relative importance of each one. Then you can plot that importance score vs. a second dimension such as satisfaction, willingness to pay, or customer size. This will produce a strategy map for the features on 2 dimensions.

Figure 4 shows a disguised example using **MaxDiff + Scale Ratings** for product features (from Bahna & Chapman, 2018). This maps *competitive perception* of a brand's features, derived from a rating scale, on the Y axis, vs. the importance of those features from a MaxDiff exercise on the X axis.

Figure 4: Plotting a Competitive Perceptions Score (Y Axis) vs. MaxDiff Importance Scores (X Axis)



Conclusion

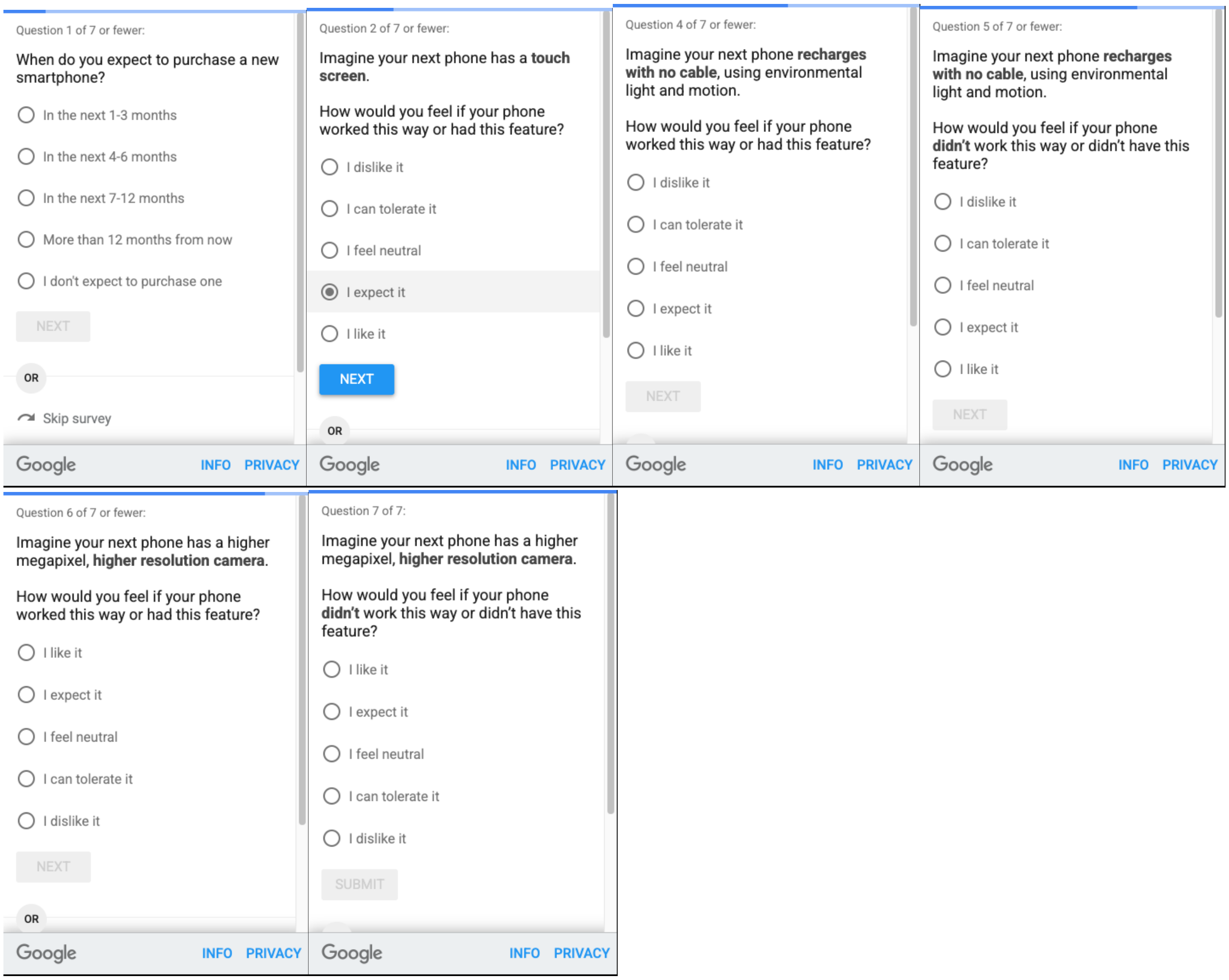
The Kano method has become popular because it attempts to answer important questions about product attractiveness. However, it appears that the theory is questionable, the commonly used items are unclear, and the item response scale is often inappropriately regarded as ordinal. The empirical results reported here suggest that the response scale is multidimensional, that responses do not strongly align with the presumed theory, and that responses have low reliability. We suggest that more traditional Likert and MaxDiff scales may be used to achieve the benefits of a two-dimensional plot for products and features, with greater validity and reliability.

Chris Chapman

Mario Callegaro

APPENDIX: EXAMPLE SCREENSHOTS FOR EMPIRICAL STUDIES

We fielded 7 versions (e.g., changing item order). Item and scale wording were identical on all.



Question 1 of 7 or fewer:

When do you expect to purchase a new smartphone?

○ In the next 1-3 months
○ In the next 4-6 months
○ In the next 7-12 months
○ More than 12 months from now
○ I don't expect to purchase one

Question 2 of 7 or fewer:

Imagine your next phone has a **touch screen.**

How would you feel if your phone worked this way or had this feature?

○ I dislike it
○ I can tolerate it
○ I feel neutral
● I expect it
○ I like it

Question 4 of 7 or fewer:

Imagine your next phone **recharges with no cable,** using environmental light and motion.

How would you feel if your phone worked this way or had this feature?

○ I dislike it
○ I can tolerate it
○ I feel neutral
○ I expect it
○ I like it

Question 5 of 7 or fewer:

Imagine your next phone **recharges with no cable,** using environmental light and motion.

How would you feel if your phone **didn't** work this way or didn't have this feature?

○ I dislike it
○ I can tolerate it
○ I feel neutral
○ I expect it
○ I like it

Question 6 of 7 or fewer:

Imagine your next phone has a higher megapixel, **higher resolution camera.**

How would you feel if your phone worked this way or had this feature?

○ I like it
○ I expect it
○ I feel neutral
○ I can tolerate it
○ I dislike it

Question 7 of 7:

Imagine your next phone has a higher megapixel, **higher resolution camera.**

How would you feel if your phone **didn't** work this way or didn't have this feature?

○ I like it
○ I expect it
○ I feel neutral
○ I can tolerate it
○ I dislike it

REFERENCES

Bahna, E., and Chapman, CN.. (2018). Constructed, Adaptive MaxDiff. *Proc. 2018 Sawtooth Software Conference.*

Bolton, K., & Brace, I. (2022). Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research (5th ed.). Kogan Page.

Chapman, C.N. (2013). 9 things clients get wrong about conjoint analysis. In B. Orme, ed. (2013). *Proc. 2013 Sawtooth Software Conference*.

J Hartmann and M Lebherz (2016). Literature Review of the Kano Model: Development Over Time (1984–2016). Whitepaper, Halmstad University.

Herzberg, F.; Mausner, B.; Snyderman, B. B. (1959). *The Motivation to Work* (2nd ed.) New York: John Wiley.

Kano, N., Seraku, N., Takahashi, F. and Tsuji, S. (1984) Attractive Quality and Must-Be Quality. *Journal of the Japanese Society for Quality Control*, 41, 39–48.

Krosnick, J. A., & Presser, S. (2010). Question and Questionnaire Design. In P. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd ed., pp. 263–313). Emerald Group Publishing.

Lavrakas, P. J. (2008). Mutually exclusive. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (Vol. 2). Sage. https://dx.doi.org/10.4135/9781412963947.n312

Martin Löfgren & Lars Witell (2008) Two Decades of Using Kano's Theory of Attractive Quality: A Literature Review, *Quality Management Journal*, 15:1, 59–75.

J Mikulić (2007). The Kano Model: A Review of its Application in Marketing Research from 1984 to 2006. Online, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.568.3350&rep=rep1&type=pdf

Raiche G, and Magis D (2020). nFactors: Parallel Analysis and Other Non Graphical Solutions to the Cattell Scree Test. R package version 2.4.1. https://CRAN.R-project.org/package=nFactors

L Witell, M Löfgren, M, and J Dahlgaard. (2013). Theory of attractive quality and the Kano methodology—the past, the present, and the future. *Total Quality Management & Business Excellence*.

D Zacarias (2015). *The Complete Guide to the Kano Model*. Online, https://www.career.pm/briefings/kano-model