

A Critique of Outcome-Driven Innovation[®]

By Gerry Katz, Executive Vice President, Applied Marketing Science, Inc.

Anyone in the field of innovation knows that, for every rule, there are a dozen exceptions. And despite the fact that most practitioners clearly advocate for their favored methods and tools, all would readily agree that there is no “one right way” that predictably guarantees success. But now, we find a noteworthy exception.

In his highly publicized book, *What Customers Want*¹, Strategyn CEO Anthony Ulwick argues that it is time to “silence the voice of the customer” in favor of his firm’s approach, which it has branded as *Outcome-Driven Innovation*^{®2}. Increasingly, our clients, prospects, and other fellow Voice of the Customer (VOC) practitioners have asked our opinion of Strategyn’s approach and how it differs from the well-established VOC methodologies practiced by AMS and others in the product development profession for at least twenty years.

Ulwick deserves commendation for contributing to the canon of new product development (NPD) and the profession’s ongoing discussion of best practices. He presents his ideas clearly and logically and then offers Strategyn’s recommended methodology in a straightforward step-by-step manner. For a newcomer to VOC, his methodology is a clear step up from the “flying blind” approach to customer needs gathering that regrettably remains the norm at too many companies. Experienced VOC practitioners, however, will see that much of his argument depends on a naïve definition of VOC and its use at companies the world over. Furthermore, longtime market researchers and product developers will be justifiably troubled by his methods, which dangerously test the limits of generally accepted good practice and raise serious questions about the integrity of the results. Product developers drawn to Ulwick’s approach must therefore thoroughly consider its inherent weaknesses before adopting this seemingly new and innovative method.

IS OUTCOME-DRIVEN INNOVATION[®] SUBSTANTIVELY DIFFERENT FROM VOICE OF THE CUSTOMER?

We are often asked, “Is Outcome-Driven Innovation really any different from Voice of the Customer?”, or more critically, “Are these not the same concepts that academics and VOC practitioners have been teaching and using for many years?” Ulwick claims that his methodology is a significant break from traditional Voice of the Customer, going so far as to declare VOC effectively obsolete. But his claim rests on a straw man built around a flawed definition of the term. He says VOC fails because it relies on *customers* to articulate the necessary features, solutions, and technical specifications. If we accept his naïve definition of VOC, we too would surely view his approach as radically different. It is not. As Jeffrey Pinegar writes in his review of *What Customers Want* for the *Journal of Product Innovation Management* (JPIM):

¹ Ulwick, Anthony W. 2005. *What Customers Want: Using Outcome-Driven Innovation to Create Breakthrough Products and Services*. New York, NY: The McGraw-Hill Companies, Inc.

² “Outcome-Driven Innovation” is a registered trademark of Strategyn, Inc.

“In proposing his method, Ulwick disparages several other techniques; nevertheless, the reader should remember that techniques are often erroneously labeled as flawed when in fact they are misapplied or poorly executed.”³

It is no surprise then that Voice of the Customer—a long-time “best practice” that has been adapted to almost every industry imaginable and over which no individual or firm can claim sole ownership—has become diluted in definition, and indeed, misapplied and poorly executed. When VOC “fails,” it is often because the team has the wrong understanding of the term, assuming that VOC means simply asking customers *what they want* outright. Rarely does this provide much insight; customers often have neither the expertise needed nor the product knowledge or technological wherewithal to be truly creative. When asked directly, we’ve observed that customers simply regurgitate what they perceive to be the best solutions currently available. Companies following these customers’ advice should not be surprised when their latest “innovation” turns out to be a me-too product and not the breakthrough they had hoped for.

So Ulwick is right to condemn this approach, which he cleverly calls the *literal* voice of the customer, but he is most certainly not the first to do so. That road has been well worn by many academics and practitioners ahead of him. To wit:

“A customer need is a description, in the customer’s own words, of the benefit to be fulfilled by the product or service ... Note that the customer need is not a solution, say a particular type of monitor (VGA, Super VGA, XGA, Megapixel, etc.) nor a physical measurement (number of noticeable breaks in the line), but rather a detailed description of how the customer wants images to appear ... If the product development team focuses too early on solutions, they might miss creative opportunities.”⁴

“Definition [of a customer need]: A sentence that describes from the customer’s (or consumer’s) vantage point the need / issue / problem that needs to be solved or solved more effectively.”⁵

“Customer needs are the problems that a product or service solves and the function it performs. They describe what products let you do, not how they let you do it ... In general, needs and problems are fairly stable, they change only slowly, if at all, over time. Features deliver the solutions to people’s problems. Features are the ways in which products function—a portable PC delivers a partial solution to being able to work wherever I want.”⁶

“In most cases the customer can only tell you what task they are trying to accomplish and what obstacles they have encountered. The goal ... is to make certain that the design engineers who are responsible for developing your technology are intimately familiar with the customer’s task demands.”⁷

“A requirement is a statement of a problem as opposed to ‘the product must do this.’ ... What outcome do customers desire that they cannot now achieve?”⁸

³ Pinegar, J.S., *Journal of Product Innovation Management*, 2006:23, p.464-466.

⁴ Griffin, Abbie, and John R. Hauser. 1993. “The Voice of the Customer.” *Marketing Science* 12, 1, Winter: 1-27.

⁵ Brodie, Christina H. 2004. Integrating a Requirements Process into New Product Development. In *The PDMA Toolbook: for New Product Development*, ed. P. Belliveau, A. Griffin, and S. Somermeyer, 331-352. Hoboken, New Jersey: John Wiley & Sons, Inc.

⁶ Griffin, Abbie. 2005. Obtaining Customer Needs for Product Development. In *The PDMA Handbook of New Product Development*, ed. K. Kahn, G. Castellion, and A. Griffin, 211-227. Hoboken, New Jersey: John Wiley & Sons, Inc.

⁷ McQuarrie, Edward F. 1998. *Customer Visits*. Thousand Oaks, CA: Sage Publications.

“Customers find themselves needing to get things done. When customers find that they need to get a job done, they hire products or services to do the job. This means that marketers need to understand the jobs that arise in customers’ lives for which their products might be hired ... The job is the fundamental problem a customer needs to resolve in a given situation.”⁹

“It is all too tempting to focus immediately and directly on product specifications. Thus, you might be inclined to ask, ‘How many megabytes do you need this instrument to transfer per second?’ Instead, you would be better off beginning with an analysis of task demands. What kinds of data are being transferred? Where do the data come from, and where do the data go? Who produces the data and who consumes it? In other words, what business purpose does transferring the data serve, and what are the criteria for success that apply to that business purpose?

The point here is that the customer’s business purpose is fundamental, not the product specification. The product will be purchased if it satisfies a business purpose, otherwise not. Hence, the most important thing is to understand the business purpose—the task the product supports—and how this purpose articulates with product functionality ... a very typical flaw is asking too many product-focused and not enough customer-focused questions. A focus on task demands, first, and product specifications, second, is one concrete way in which you give meaning to the quest for a market-focused business strategy.”¹⁰

In all of these cases, the authors make a clear distinction between *needs* and the *solutions* to those needs. Ulwick adds the term *desired outcomes* to this lexicon, a useful description to be sure, just as Christensen has popularized the term *jobs*. But neither of these is conceptually any different from the other terms that have been in use since at least the mid-1980s: *wants*, *needs*, *requirements*, *benefits*, *problems*, *tasks* that the customer is trying to accomplish, and *jobs* which the customer wishes to get done.

Does Data Collection Method Matter?

Strategyn also claims that focus groups, one-on-one interviews, and ethnography work equally well at eliciting customer needs. Academics and practitioners disagree. In their landmark 1993 paper¹¹ (just selected as one of the 20 most influential papers in the history of Marketing Science by INFORMS – the Institute for Operations Research and Management Science), Abbie Griffin and John Hauser empirically compared focus groups and one-on-one interviews. They proved that one-on-one interviews are significantly more effective than focus groups for gathering customer needs. For one, they showed that one-on-one interviews achieve nearly the same result as focus groups—identification of virtually 100% of the customer needs in a category—but at significantly lower field expense. Moreover, they observed that one-on-one interviews allow the interviewer to explore tangents with individual customers, which often yields valuable new information, without alienating other participants.

Regarding ethnography, i.e. the use of observational techniques to uncover needs and insights, Ulwick’s argument is even more controversial. In VOC, ethnography plays a critical role in

⁸ Mello, Sheila. 2002. *Customer-Centric Product Definition: The Key to Great Product Development*. Boston, MA: PDC Professional Publishing.

⁹ Christensen, Clayton M. 2007. “Finding the Right Job for Your Product.” *MIT Sloan Management Review* 48, 3, Spring: 38-47.

¹⁰ McQuarrie, Edward F. 1998. *Customer Visits*. Thousand Oaks, CA: Sage Publications.

¹¹ Griffin, Abbie, and John R. Hauser. 1993. “The Voice of the Customer.” *Marketing Science* 12, 1, Winter: 1-27.

uncovering *latent, hidden, unarticulated, or difficult to articulate* needs. He argues, instead, that such needs do not exist and that simply by using his prescribed questioning methodology, *all* needs will be articulated (although he never quite spells out what that questioning methodology is). Our experience flatly contradicts this point of view. We would agree that a highly-skilled interviewer can elicit some needs that might be “hidden” to a layperson, but ethnographic techniques surely increase the likelihood of finding them. For instance, it is well known that the advent of cupholders in automobiles happened as a result of observation, as customers struggled to hold their beverage cups on the center console or between their legs. Likewise, in a VOC study on dialysis equipment, we observed dozens of nurses searching for a flat surface on which to record their notes, sometimes walking clear across the clinic floor to find one. Yet during the interviews, not one of them actually articulated the need for a surface on which to write; but once we observed its existence, it was quite easy to probe respondents further. Without observation, however, this important unspoken need would still be hidden.

Mode of Questioning

To uncover wants, needs, jobs, and desired outcomes research has shown that the most useful method is to discuss past *experiences* with these types of products or services in trying to accomplish the desired tasks.

“Use indirect questioning, rather than direct questions. Thus, rather than asking customers “What do you want” directly, ... VOC indirectly discovers wants and needs by walking customers through the ways they currently obtain or acquire and use products and services to fulfill particular needs.”¹²

Experienced market researchers use a particular questioning technique that recognizes the distinction between *needs* and *solutions*. Whenever a customer suggests a feature, solution, or technical specification, the interviewer must probe beyond this original comment to understand the underlying need that the product must address.

“Many customers will offer a solution, an engineering characteristic, or a laboratory-measurable technical specification that they think does a good job of addressing their need. For instance, ‘The exterior wall should be an alloy of aluminum and titanium.’ Whenever this happens, a good interviewing technique is to ask them *why* they think that would be a good solution. This often provokes them to state the real underlying need ...”¹³

The goal of a Voice of the Customer interview is to get to the details that lie beneath these potential features and solutions. Any trained VOC interviewer knows to ask the customer what they are trying to accomplish, or what the job is that they are trying to get done. Ulwick implies that his ideas are new and revolutionary, and that they have transformed New Product Development! Clearly, there is ample evidence to the contrary. This is merely old wine in new bottles.

¹² Griffin, Abbie. 2005. Obtaining Customer Needs for Product Development. In *The PDMA Handbook of New Product Development*, ed. K. Kahn, G. Castellion, and A. Griffin, 211-227. Hoboken, New Jersey: John Wiley & Sons, Inc.

¹³ Katz, Gerald M. 2004. The Voice of the Customer. In *The PDMA Toolbox: for New Product Development*, ed. P. Belliveau, A. Griffin, and S. Somermeyer, 167-199. Hoboken, New Jersey: John Wiley & Sons, Inc.

Level of Detail

Ulwick notes that, in most product categories, 100 or more detailed needs are common – an observation borne out by our experience as well. However, it has long been accepted that this is simply too much detail for a product development team to deal with. Thus, virtually every author and practitioner in the field recommends the concept of creating an *affinity diagram*, a simple way to organize the needs into a hierarchy based on some logical method of organization, and thus reducing the complexity of dealing with 100 needs to a more workable 15-25 groups of highly related needs.

Strategyn stands apparently alone in believing that it is acceptable—desirable even—to skip over this step, and instead advocates that companies and customers deal with all 100 or more needs (*outcomes*, in their terminology) individually. This approach presents a serious problem from the standpoint of data collection, in that it requires survey respondents to prioritize *all* of the desired outcomes on a minimum of two dimensions—*importance* as well as the *performance* of current alternatives. Ask the respondent to rate the performance of two or even three separate products or services—essential for product-to-product comparisons—and it is quite possible that each respondent must answer three or four hundred questions!

Respondent fatigue is a real concern in survey research, and responsible market researchers would surely recognize Ulwick’s approach as unreasonably tedious for the respondent. Long and repetitive surveys lead respondents to drop out before completing the questionnaire, potentially driving down effective incidence (and consequently driving up cost), and possibly introducing “survivorship bias” by only considering the responses of those who complete the survey. Worse, long surveys encourage “gaming” such as “straight-lining,” or “zig-zagging” where respondents stop thinking carefully and just check the boxes willy-nilly to reach the end and collect the incentive.

Client fatigue may be an even more troubling aspect of Strategyn’s approach. Several companies who have used Strategyn’s methodology have told us that they nearly “drowned” in detail when it came to analyzing and making sense of the data. Senior managers are not interested in excruciating detail, but rather in a meaningful and actionable summary of the data and its implications. The use of an affinity diagram allows management to synthesize the results of their VOC, helping them to make intelligent R&D and engineering investment decisions, and moving more quickly toward solutions.

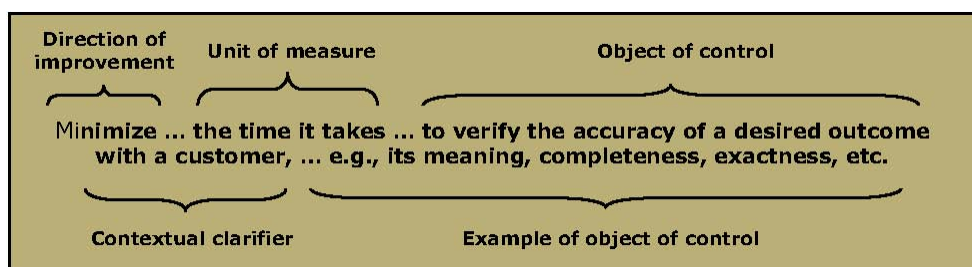
Creating an affinity diagram is a clear case of “less is more,” both for the respondent and for the interpretability of the data. Since most solutions tend to impact most if not all of the needs in a given cluster, to keep them disaggregated creates a bit of a dilemma. What does one do when several needs within the same cluster receive radically different prioritization scores? Some form of a composite score, as is the case with an affinity diagram, is actually advantageous here.

Most people in the field of new product development believe that something like 15 to 25 constructs is the right level of detail for the optimal organization and focus of new product development activities. To be sure, we have found it to be extremely worthwhile to maintain the list of detailed needs that make up each cluster on the affinity diagram. (And when additional

detail about a particular cluster is needed, one can always conduct a separate survey to “drill down” on just those individual needs.) But when a final report presents ratings of 100 or more items, the product developer is likely to end up with too much data and not enough useful information.

Structure of Needs Statements

Outcome-Driven Innovation requires that customer needs (desired outcomes) be expressed using a specific syntax. According to *What Customers Want*, each phrase must contain a direction of improvement (i.e. the words “maximize” or “minimize”, “increase” or “decrease”), a unit of measurement, and an object of control, along with a contextual clarifier and one or more examples of the object of control, if necessary. For example:¹⁴:



Ulwick and Bettencourt are not the first to tread this road; other VOC practitioners have argued that customer needs be expressed in a defined way. Some practitioners insist that each need begin with, “I need...” or “A [product] that lets me...” All follow the same rationale; standardizing the format of each need is thought necessary to remove ambiguity for both the customer and the product developer.

In our early work on affinitization and prioritization of needs with customers, however, we found that this type of repetitive language was more of a hindrance than an advantage. Once again, the culprit is respondent fatigue. We observed that respondents were less likely to complete the affinitization task, and even among those who did, the data looked less and less logical over time. On closer examination, it was clear that respondents began to take the exercise less seriously, giving little or even no thought the further along they went. Moreover, we were only considering 15-25 needs clusters. Adding Strategyn’s requirement that customers prioritize *all* 100+ needs, each starting with the word “minimize” or “maximize”, as opposed to the 15-25 needs clusters, and the task becomes truly overwhelming.

Is a Fixed Syntax Really a Benefit?

Strategyn’s insistence on a uniform syntax has a certain appeal—removing variation from a process always appears to be more scientific; however, in this case it creates several new

¹⁴ Ulwick, Anthony W, and Lance A. Bettencourt. 2008. “What is Outcome-driven Innovation?” *MIT Sloan Management Review* 49, 3, Spring: 62-68.

problems. First, customers do not often say “*minimize*” or “*maximize*” when stating their needs. As a result, the interviewer must either train the customer to phrase responses this way, or alter the customer’s words, force-fitting them into the required syntax. The former risks annoying or even insulting the customer, while the latter violates a best practice in VOC: to preserve the customers’ own words as much as possible, so as not to alter their meaning or inadvertently read a different meaning into what the customer has actually said.

This principle is fundamental to VOC, and with good reason. Before VOC became *de rigueur* in new product development, engineering teams routinely took to brainstorming, imagining, or simply guessing what their customers wanted. Products were thus built upon the Voice of the *Company*, or worse, the Voice of the *Engineer*, not the Voice of the *Customer*. And despite the best of intentions, the consequences (predictably) were often disastrous. Over time, academics and practitioners came to realize that preserving the customers’ language was the surest way to keep the Voice of the *Company* at bay. Strategyn’s approach, which requires the interviewer to manipulate the customer’s words into a customer-defined metric, runs counter to the prevailing wisdom of VOC, injecting the Voice of the *Market Researcher* into the mix and risking a return to the age-old Voice of the *Company* problem.

Several companies who have used the Strategyn approach have told us that an excessive amount of time is spent re-stating customer phrases into the desired syntax, often resulting in phrases that are overly wordy and even less clear than what the customer actually said. Sometimes it even results in phrases with awkward double negatives! In hindsight, they felt that many needs simply did not fit the rigidity of the desired outcomes syntax. To illustrate, in a VOC study we performed years ago for a cinema company, we regularly heard two straightforward needs that customers expressed in just a few words. In both cases, applying the ODI syntax complicates matters; the resulting need is wordier and much less clear:

Example A:

What the customer said:

- “*No sticky floors*”

How this might be expressed using the ODI syntax:

- *Minimize the amount of stickiness on the floor of the theater from spilled soft drinks*
- *Maximize the cleanliness of the floor in the theater, e.g. no stickiness from spilled soft drinks*

Example B:

What the customer said:

- “*Clean restrooms*”

How this might be expressed using the ODI syntax:

- *Maximize the cleanliness of the restrooms in the theater, e.g. clean floors, toilets, sinks, etc.*
- *Minimize the amount of messiness in the theater restrooms, e.g. no paper on the floor, no overflowing trash receptacles, etc.*

In each case, neither of the new phrases are any clearer than what the customer actually said. Instead, the ODI syntax yields awkward statements that are several degrees removed from the customers' own statements.

Another problem arises when needs do not have a clear direction of improvement, but rather, an ideal target, such that increasing or decreasing the measure would be worse. For instance, most people desire a beverage with exactly the “right” amount of sweetness or a wine with the “right” amount of oakiness—neither too much nor too little. Maximizing or minimizing such a measure would be contrary to the customer's expressed need. To illustrate, consider an example from a recent *pro bono* study conducted by AMS for the Product Development and Management Association (PDMA), on what makes a good professional society:

Example C:

What the customer said:

- *“I always know when it's time to renew my membership”*

How this might be expressed using the ODI syntax:

- *Minimize the likelihood that a member forgets when it is time to renew his/her membership.*

Both statements will probably result in the exact same solution, but the former says it more clearly, and is more consistent with what the customer actually said.

Strategyn further justifies its rules for syntax claiming that, without them, the prioritization ratings can change dramatically. But they provide no published scientific data to support this assertion. This claim simply lacks face validity; there is no logical reason that someone would give a significantly different importance or satisfaction rating to *no sticky floors* versus *minimize the amount of stickiness on the floor of the theater from spilled soft drinks*? Thus, we decided to do our own experiment. Not only did we find no evidence to support Strategyn's claim, but we also observed a number of other flaws in their approach that should give pause to any researcher.

An Experiment: Comparing Customer Language with the Strategyn Syntax

As a test case, we repeated the quantitative phase of our study on movie theaters – a real-world success story in that it resulted in the creation of “stadium seating” and the more extensive, premium food and beverages found in today’s theaters. In this example, we started with 75 customer needs statements, all of which were expressed in words and phrases customers actually used in one-on-one interviews. We then re-cast all 75 phrases using Strategyn’s suggested syntax as described earlier. This was not an easy task (as we had been warned by past users of this methodology) and involved several rounds of revisions to apply the strict syntactical rules to translate the needs into “desired outcomes.” Even so, many of the needs did not lend themselves easily to this syntax, resulting at times in complicated phraseology or double negatives.¹⁵

Next, we created a typical web-based prioritization survey. After a few screening questions to qualify respondents (i.e. those who attend movies regularly), participants were invited to complete a basic questionnaire consisting of three parts:

1. Rating the *importance* of each of the 75 needs using a 100-point scale.
2. Rating the *performance* of a specific, recently-attended theater in meeting each of the 75 needs using a 10-point scale.
3. Answering a set of basic demographic questions to make sure that the sample was fairly balanced and reflective of the general population.

We then fielded the survey using Survey Sampling’s consumer web panel, with of the aim of achieving about 300 completed questionnaires. People were randomly assigned to one of two cells at the outset. Both cells completed identical surveys *except* for the wording of the needs: about half received the customer-worded needs (hereafter referred to as the “Customer cell”) and half received the Strategyn syntax-worded needs (hereafter referred to as the “Strategyn cell”).

Our hypothesis was simply that there would be little, if any, significant difference in the ratings between the two cells. And indeed, a comparison of the data yielded little evidence to reject this hypothesis. Yet surprisingly, we observed several other disturbing characteristics in the test data that call into question the quality and effectiveness of the Strategyn approach.

Data Cleaning

538 respondents passed the screener – 248 for the Customer cell and 290 for the Strategyn cell – and thus, began the survey. But as with any survey, some did not finish. While there could be many factors that might influence the dropout rate, in our experience, fatigue—e.g. excessive time required or the repetitiveness of the questionnaire—tops the list. What did we find?

- **Fact:** the dropout rate for the Strategyn cell was more than 40% higher than for the Customer cell.

¹⁵ Strategyn claims that it takes about one to two years of training and practice to master this syntax, which includes more than 15 “rules” that must be followed. Some have suspected this claim to be self-serving, in that it requires that users remain dependent on them for at least this long. Despite our lack of experience with the syntax, we sincerely tried to adhere to their rules as faithfully as possible.

	Customer Cell		Strategyn Cell	
	<i>n</i>	%	<i>n</i>	%
Dropout Rate	55	22.2%	92	31.7%

This left us with 391 people who actually completed the survey – 193 for the Customer cell and 198 for the Strategyn cell. We then performed the following industry-accepted data cleaning procedures on both data sets:

1. We examined the time each respondent spent completing the questionnaire, and we removed respondents who took too little or too much time. The former—dubbed “speeders”—most likely completed the questionnaire randomly and hurriedly merely to obtain the incentive, with little thought to the scores they gave. The latter, on the other hand, probably completed the questionnaire over several separate sessions, risking discontinuity in their ratings.
2. We checked for reasonable variation in the ratings, and removed “straight-liners” – which are generally considered to be fraudulent responses. This was done separately for both the importance ratings and the performance ratings.
3. We looked for “effective quitters”—respondents who began the survey giving what appeared to be reasonable ratings, but who then skipped large blocks of questions part-way through (by answering “Not Applicable”), essentially abandoning the survey. Likely causes are unreasonable survey length, repetitive questions, or confusion over the exact task at hand.

The typical survey completion time was between 15 and 20 minutes. Therefore, we set thresholds for exclusion at less than 5 minutes or more than 60 minutes. This excluded 29 respondents, 17 from the customer cell and 12 from the Strategyn cell. We then compared the average time for completion among the remaining respondents and made the following observations:

- **Fact:** The Strategyn cell respondents took 21% longer to complete the survey than the Customer cell respondents. (This might also explain the much higher dropout rate among the Strategyn cell respondents.)

	Customer cell	Strategyn cell
Time to complete (min:sec)	15:15	18:30

Next, with regard to variation, we looked for anyone whose importance ratings had a standard deviation less than 5.0 (on a 100-point scale) or whose performance ratings had a standard deviation less than 0.5 (on a 10-point scale). This resulted in the removal of 49 more respondents – 19 from the Customer cell and 30 from the Strategyn cell. What is going on here?

- **Fact:** The Strategyn cell respondents were about 50% more likely than the Customer cell respondents to commit one or both of these types of “fraudulent” responses:

	Customer cell (n = 176)		Strategyn cell (n = 186)	
	n	%	n	%
Importance: (Std. Dev. < 5.0)	1	0.6%	9	4.8%
Performance: (Std. Dev. < 0.5)	19	10.8%	26	14.0%
Both	1	0.6%	5	2.7%
Either	19	10.8%	30	16.1%

None of this should be surprising, as it just empirically confirms some of the problems discussed earlier in this paper: that lengthy, repetitively-worded questionnaires cause negative repercussions on data quality.

The data cleaning exercise left us with 313 usable respondents, 157 from the customer cell and 156 from the Strategyn cell.

Analysis

We began by simply calculating the mean importance and performance scores on all 75 needs and plotting them on a simple graph for both cells (in decreasing order of importance / performance based on the Customer Language cell) (Exhibit 1). Immediately, we observed something quite unexpected.

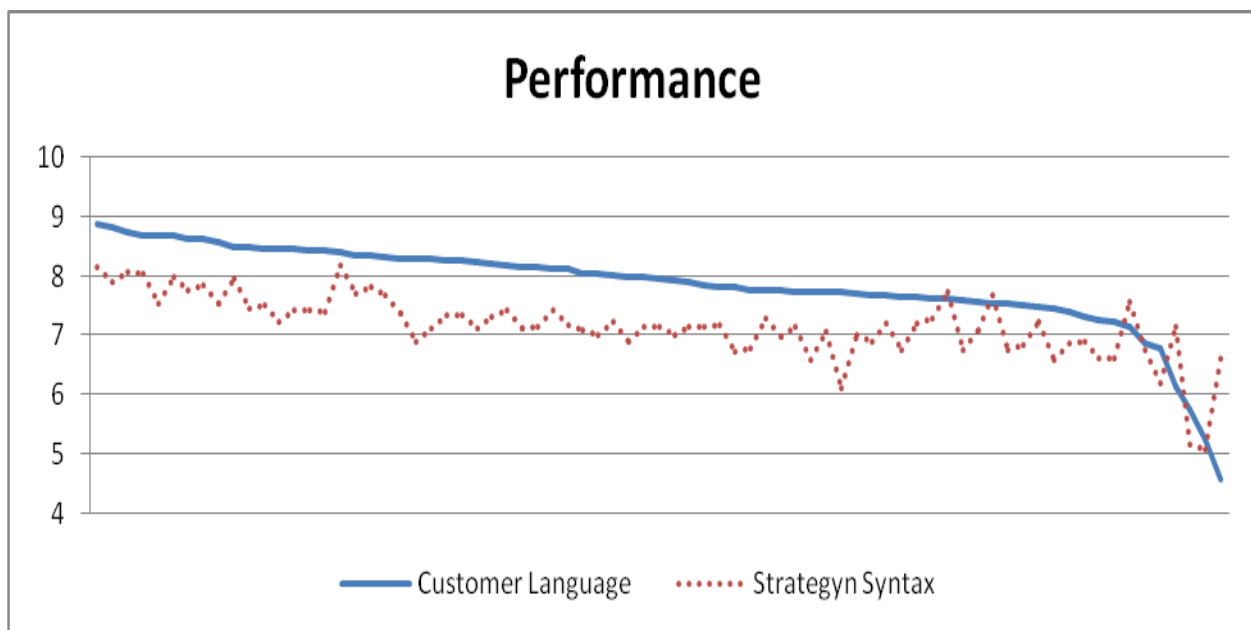
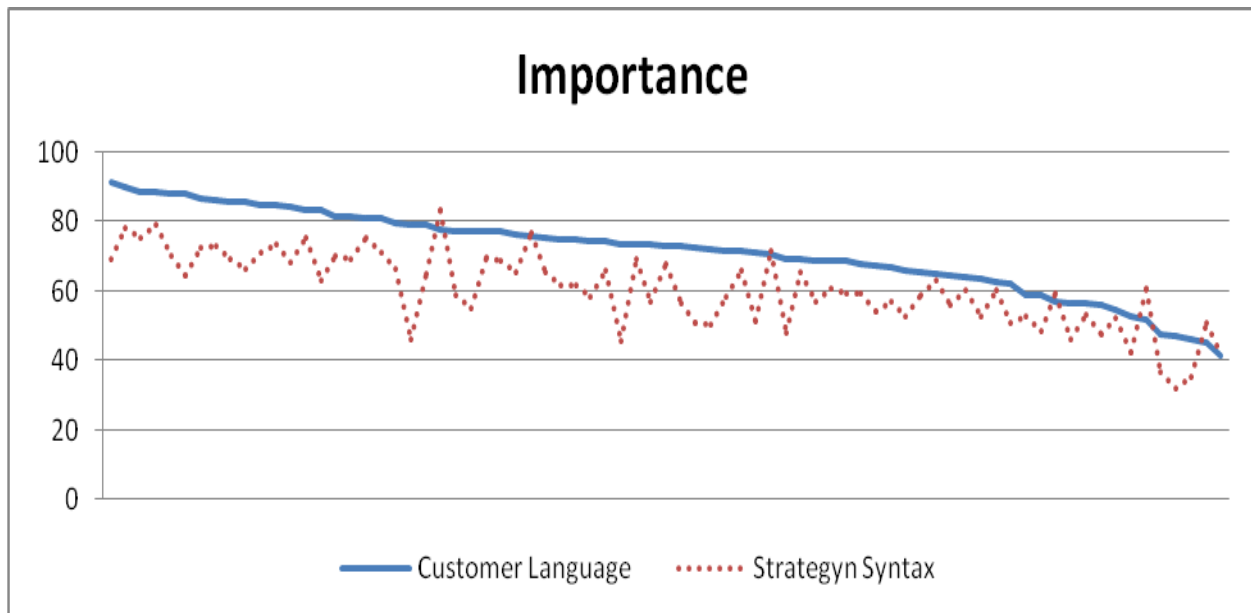


Exhibit 1

With only a few striking exceptions, the ratings move largely in lockstep, supporting our hypothesis that there is no significant difference between the two methods. There is, however, a clear and persistent positive bias to the Customer cell ratings compared to the Strategyn cell ratings. Why might this be?

One explanation could be that the Customer wording was simply more “respondent-friendly” than the Strategyn syntax which, because of the complicated and repetitive phraseology, can be

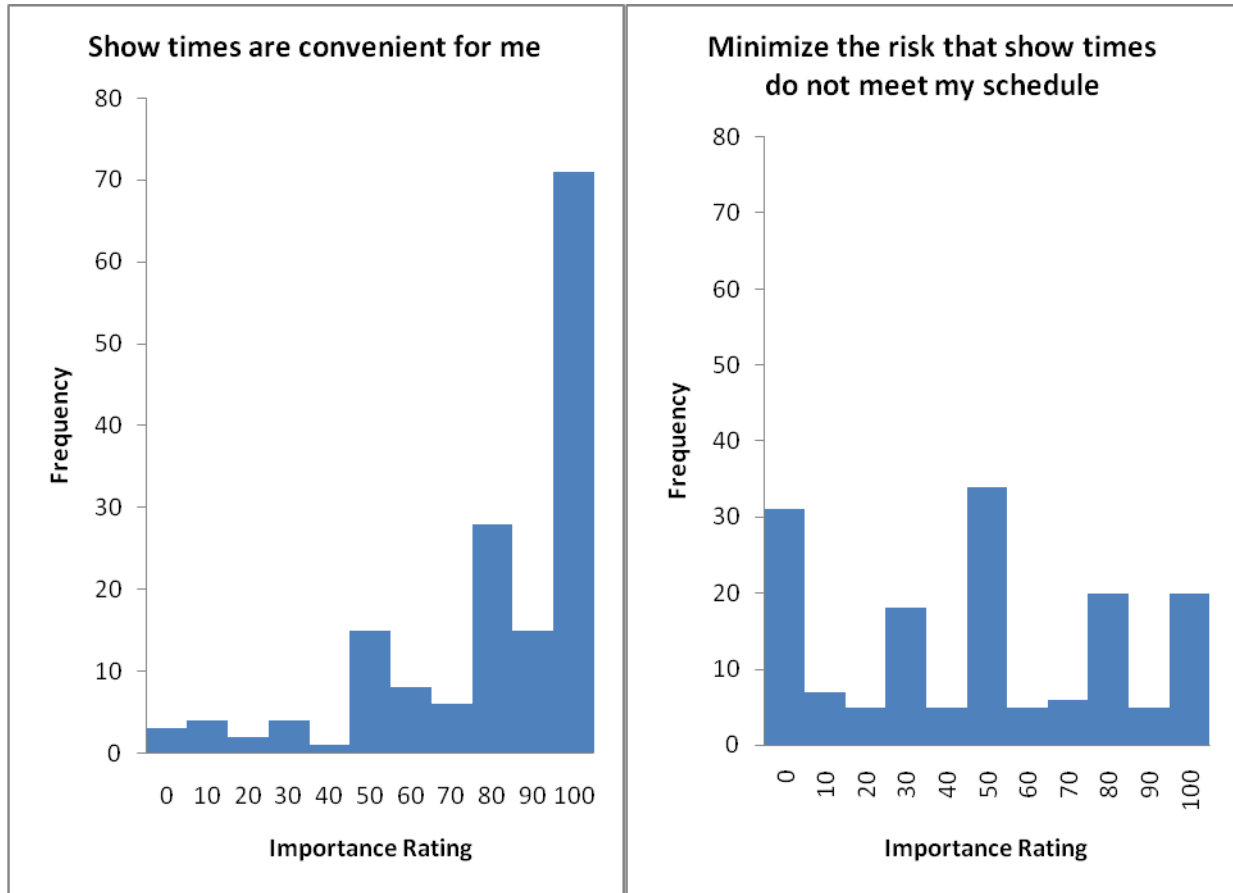
somewhat “off-putting”. While this can’t be proven, we believe it to be at least part of the explanation.

But upon further inspection, we found an even more disturbing phenomenon. Experience shows that respondents tend to skew their ratings into the upper part of any scale (more about this later). This is doubly true for ratings like importance that often have a clear, unambiguous direction of “goodness”. For instance, a need like “I am guaranteed a seat when I purchase a ticket” might be more important to some people than it is to others. But it would be illogical for someone to rate that as completely unimportant.

When we plotted the frequency distributions for each need individually and looked at the Customer cell and the Strategyn cell respondents side by side, we observed an unexpected multi-modal, “pitchfork” pattern in many of the responses using Strategyn’s syntax that did not exist when the customers’ own words were used. Consider the following:

Customer cell

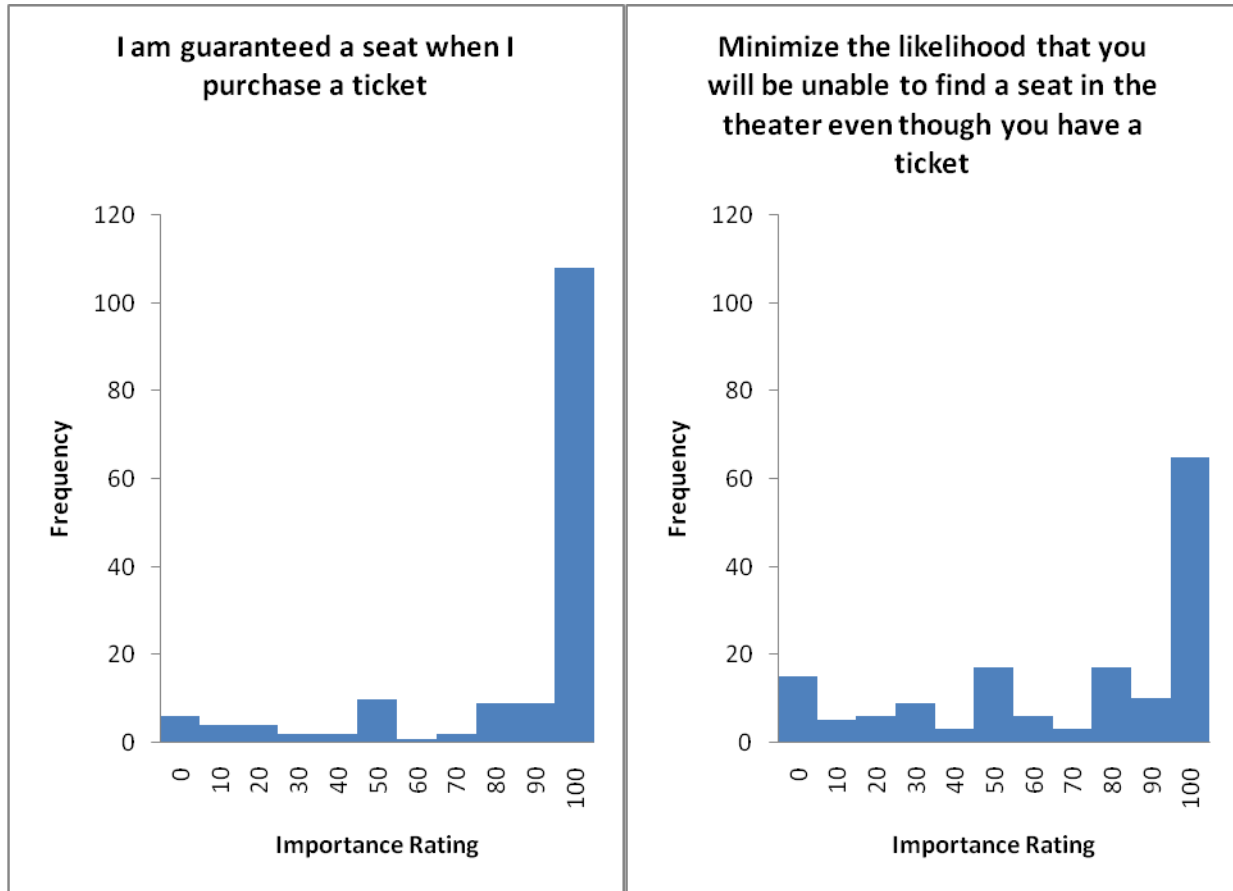
Strategyn cell



	Customer Cell	Strategyn Cell
Mean	79	46
Standard Deviation	29.07	33.47
% respondents who rated '0'	2%	20%

Customer cell

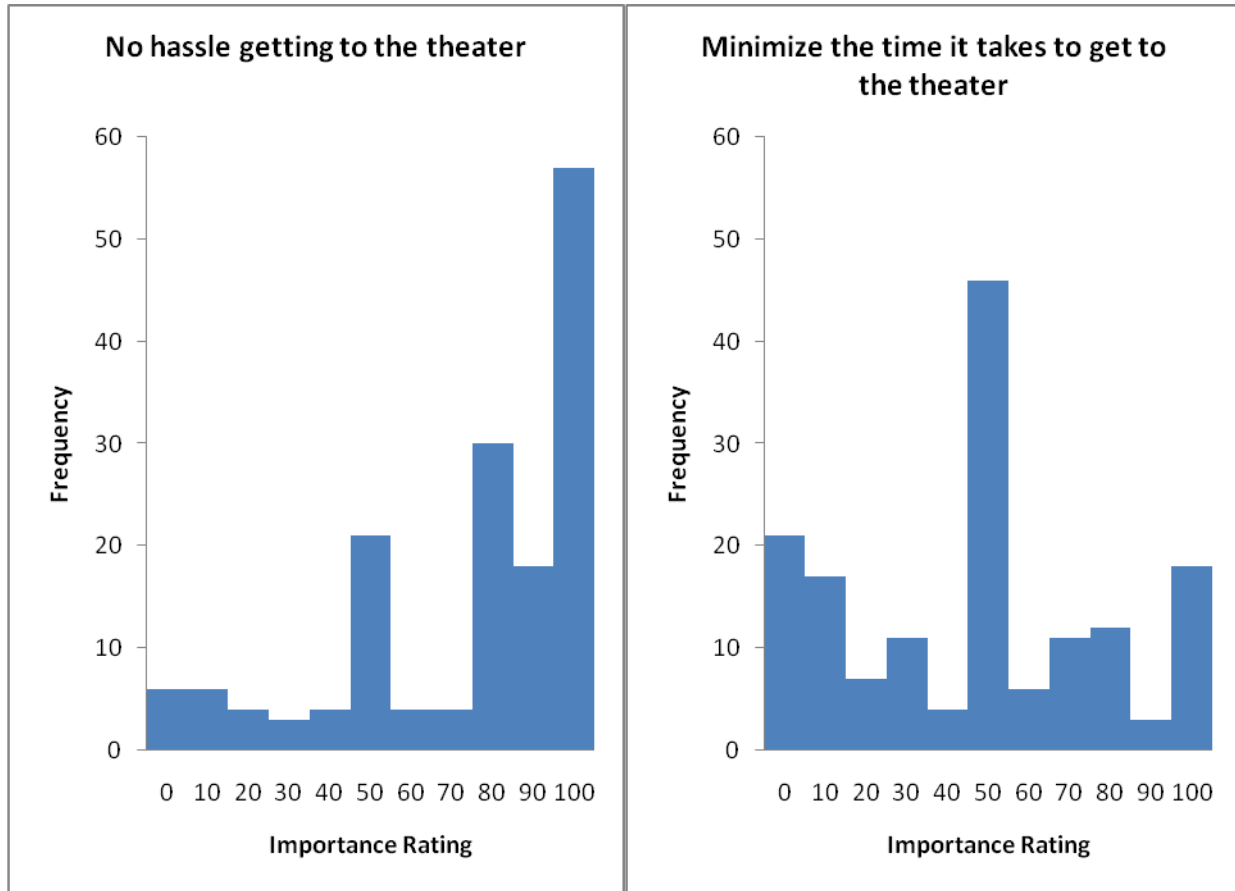
Strategyn cell



	Customer Cell	Strategyn Cell
Mean	84	68
Standard Deviation	29.35	35.62
% respondents who rated '0'	4%	10%

Customer cell

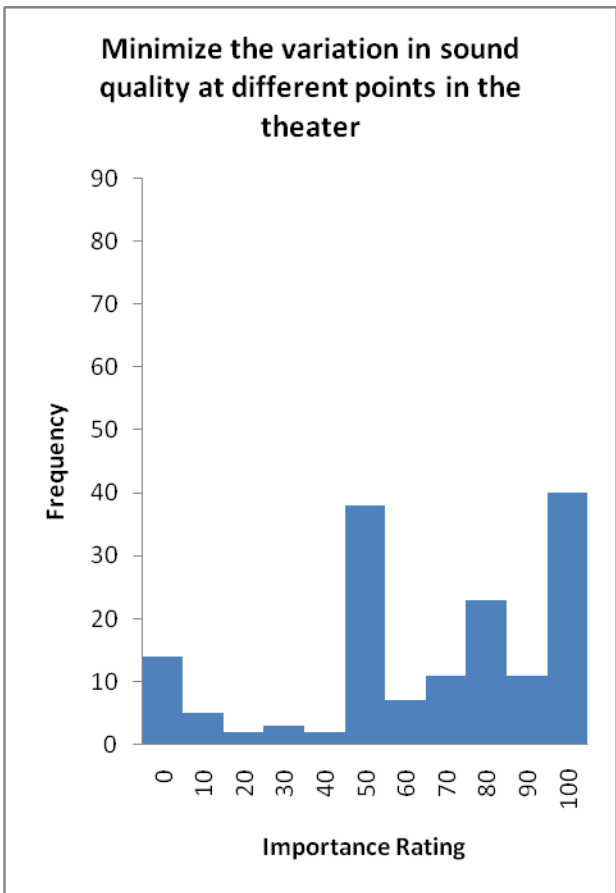
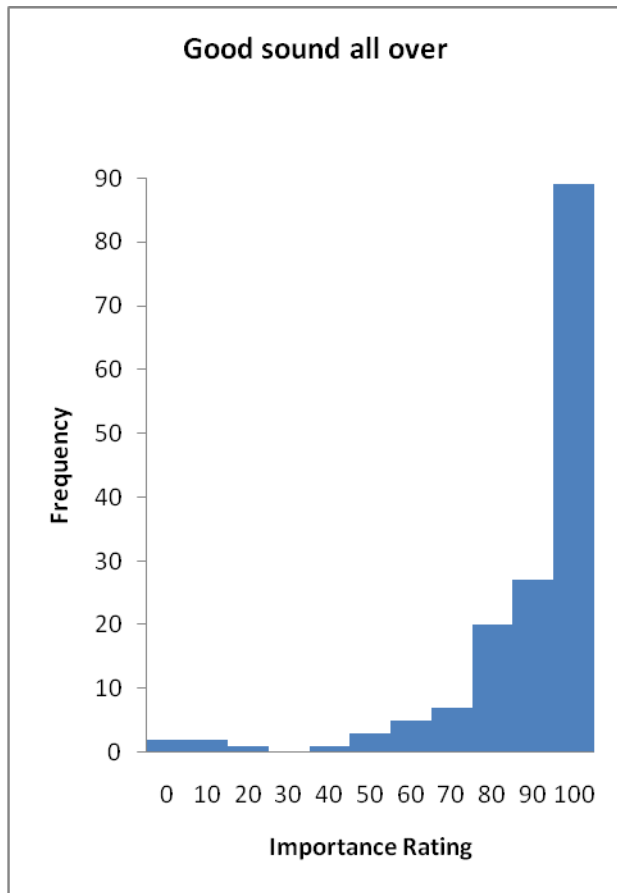
Strategyn cell



	Customer Cell	Strategyn Cell
Mean	73	46
Standard Deviation	29.72	31.51
% respondents who rated '0'	4%	14%

Customer cell

Strategyn cell



	Customer Cell	Strategyn Cell
Mean	88	64
Standard Deviation	19.73	31.49
% respondents who rated '0'	1%	9%

These are only four examples, but they provide clear evidence of respondent confusion, resulting from Strategyn's convoluted phraseology and double negatives. The multi-modal data distribution suggests that many respondents cannot clearly identify the favorable end of the scale, (which indicates greater importance). The "spike" observed in the middle of the scale further strengthens the case for confusion, indicating cases where respondents selected the middle of the scale to avoid commitment and "split the difference" rather than risk a "wrong" answer. This phenomenon occurs only rarely in the Customer cell, but it happens quite visibly on about a third of the needs in the Strategyn cell. It may also explain the consistently lower ratings when following the Strategyn syntax, as the higher frequency of low scores pulls down the average.

Adding to the problem, this experiment included only 75 needs. However, Strategyn claims that most of their studies result in between 100 and 150 needs, further exacerbating the problem. Clearly, it is safer to affinitize the needs so that respondents are only asked to rate the 15-25 resulting needs clusters.

Testing Our Hypothesis

In order to test our original hypothesis, we first had to deal with the bias, i.e. the consistently higher ratings given in the Customer Language cell, as explained above. We decided to simply calculate the grand mean of the differences between the two cells and subtract this constant from each of the Customer cell ratings to remove the positive bias and bring the two series into better alignment, as follows (Exhibit 2).

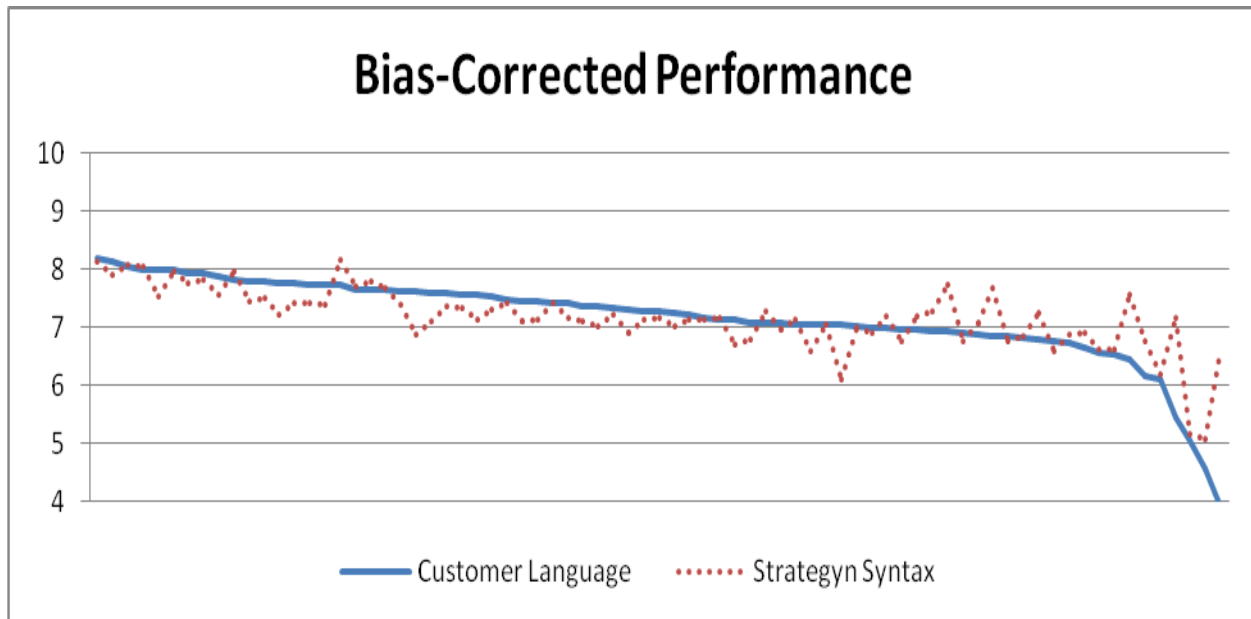
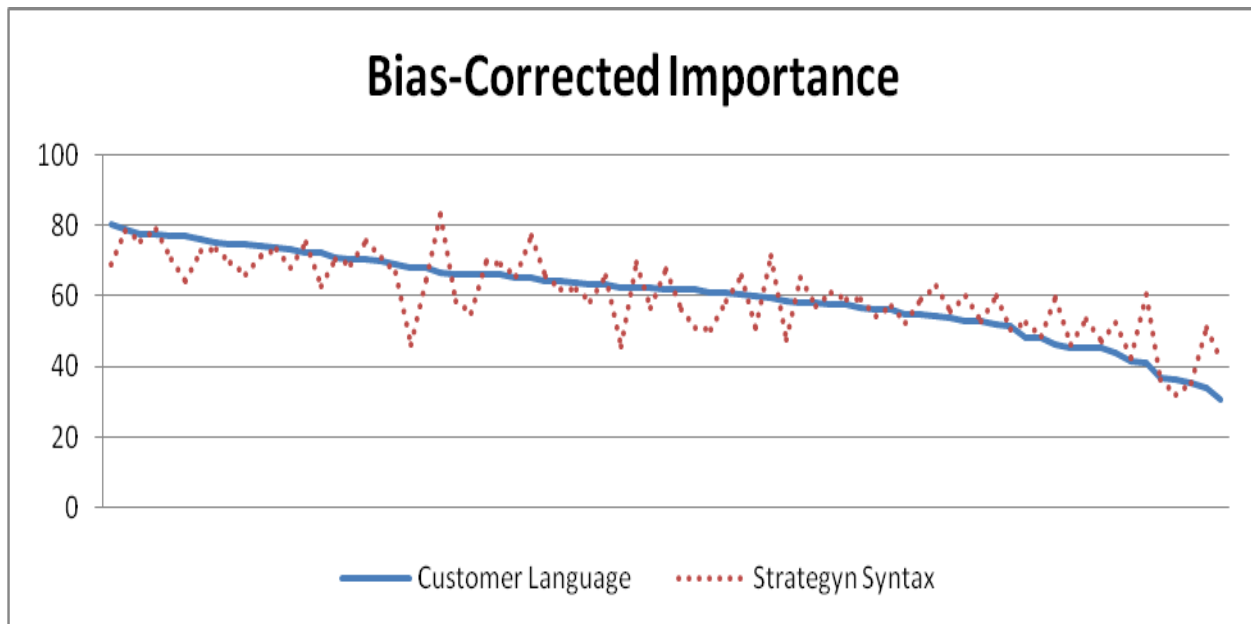


Exhibit 2

Next, we performed a t-test on each of the 75 needs for both importance and performance. With a 90% confidence level, only 10 of the 75 needs (13%) were rated significantly different on importance, and only 5 of the 75 (7%) were rated significantly different on performance. Clearly Strategyn’s justification for their syntax – that it removes variation from the prioritization ratings – is false. (Even if it were true, who is to say which wording is more “correct”, a question that is simply unanswerable.) Given all of the other phenomena observed in this experiment – the

higher dropout rate, the longer time needed to complete the survey, the higher fraud rate, and the large number of illogical ratings that probably indicate confusion, the Strategyn syntax should be viewed as highly troubling, and definitely not an advantage over the use of customer-worded needs.

Retiring the House of Quality

In a more recent monograph, Ulwick and longtime QFD gurus Richard Zultner and Rick Norman assert that it is time to “retire the House of Quality,” in favor of Outcome-Driven Innovation. The House of Quality matrix—the first and most commonly-used matrix in Quality Function Deployment (QFD)—helps translate customer needs into technical specifications, performance measures, and other metrics that developers can use to guide the development of and then evaluate potential alternative solutions. The authors argue that since Outcome-Driven Innovation gets the *customer* to articulate these specifications as measurable outcomes, the role of QFD is eliminated. (This may also explain why Strategyn insists on a specific syntax requiring a “direction of improvement” a “unit of measure,” and an “object of control,” all of which are constructs borrowed from QFD.)

However, this approach demonstrates a considerable lack of understanding of QFD, and as with VOC, Ulwick et. al. seem to choose a definition that suits their own purposes. QFD is almost always carried out as an *internal* function; customers are generally not involved, because in most cases they lack the technical knowledge to properly translate needs into product design specifications.

By way of example, imagine you lead the design team for a new generation of shaving razors. Strategyn’s approach to VOC leads us to conclude, quite correctly, that customers want to “minimize the risk of nicks and cuts when drawing the blade over skin.” Even if customers could give an exact indication of the acceptable risk of nicks and cuts, as an engineer, there are perhaps a dozen or more technical specifications that one could adjust to achieve such an outcome. You might adjust the blade’s angle of inclination, increase the blade’s sharpness, use a more rigid metal, vary the number of blades, change the distance between blades, or increase the blade’s resistance to corrosion. In fact, you might find that the optimal design comprises a combination of all of these.

Therefore, understanding that customers want to “minimize the risk of nicks and cuts” is necessary, but insufficient. Engineers must take what customers want and then translate it into engineering characteristics: concrete measures of laboratory specifications or business processes that, when optimized, result in greater customer satisfaction. The “outcomes-based” approach, with its customer-defined metrics, does nothing to address this problem.

The authors of this monograph do point out (correctly, we believe) that a full, formal QFD may be unnecessary to perform this translation process. Judging by our experience, few companies still do. Yet to claim that QFD can be “retired” simply because customers now define the metrics, or by manipulating the customer needs statements into the form of metrics is dangerously misleading. In most cases, the customer is simply not qualified to do so, and thus the scientists and engineers must still perform this task in one way or another.

The Opportunity Algorithm

Of course the real test of ODI should be the results and the value it provides to managers—in other words, does it help managers make better decisions for NPD strategy? Clearly, the goal of all VOC is to identify important, unmet customer needs. After collecting the importance and performance scores for the 100+ outcomes using a customer survey, Strategyn’s approach applies an equation—once again branded as the “Opportunity Algorithm[®]”¹⁶—which combines the scores to arrive at a single value for each need:

$$\text{Opportunity} = \text{Importance} + \max(\text{Importance} - \text{Satisfaction}, 0)$$

This algorithm is both technically and intellectually flawed. Importance and performance are two entirely separate constructs, and should not be used in the same equation. Upon first seeing this equation, MIT professor John Hauser commented:

“This measure is pseudo-scientific. It mixes units of measure. No self-respecting engineer would ever do such a thing.”

And in his JPIM review of the book¹⁷, Jeffery Pinegar observed:

“Technically, there are two problems with this formula. First, satisfaction is subtracted from importance; this is like subtracting apples from broccoli. Ulwick casually dismisses the criticism, saying that ‘it doesn’t hold up when talking about jobs, outcomes, and constraints’ (p. 47), but offers nothing to support his proposition other than anecdotes.”

In essence, Ulwick acknowledges the problem, but justifies its use by simply claiming that it works.

And indeed it does – in one respect. Strategyn’s algorithm does accomplish the objective of identifying important, unmet attributes and needs—the same goal of what has come to be known as *gap analysis*, a technique that product developers have used for years. The highest opportunity scores, analogous to the widest gaps, generally represent the key unmet needs in the marketplace, those on which product developers should target their efforts aggressively.

However, both the design of the algorithm and the scores that comprise its inputs cause it to miss the mark on some other key issues:

A. Rating Scale

Strategyn uses a basic five-point scale for both importance and performance / satisfaction. Since it is a well-known phenomenon that respondents tend to skew most of their responses to only the top part of almost any scale used, a five-point scale is likely to lack sufficient “granularity”, i.e. the ability to discriminate between small differences. As a result, most researchers use at least a 7-point scale, and many insist on a 9- or 10-point scale, or even higher.

¹⁶ The Opportunity Algorithm[®] is a registered trademark of Strategyn, Inc.

¹⁷ Pinegar, *ibid.*

B. *Top-Two Boxes*

Strategyn specifies that the inputs to be used in the algorithm are not the *average* ratings for importance and performance, but rather the *percentage* of responses that are rated either a “4” or “5” – a measure known as “top-two boxes”, which has fallen out of favor elsewhere and remains controversial. Consumer packaged-goods researchers once used a similar five-point scale to measure purchase intent; that is, respondents who said that they *definitely* intended to buy the product (a rating of “5”) and those who *probably* intended to buy the product (a rating of “4”) were lumped together. Intuitively, a “5” meant something considerably stronger than a “4,” but *top two boxes* persisted, until the BASES® model¹⁸ proved empirically that the “*definitelys*” were often five times more likely to purchase the product in reality than the “*probablys*”. Thus, considering two different products—one with 10% “definitely” and 40% “probably” and one with 20% “definitely” and 30% “probably”—both would be seen as equivalent when measured by *top two boxes*. But in truth, the latter product is far more attractive to consumers. So, combining the top two boxes has been replaced with an overall average rating by most practitioners, a method that takes these differences into account. Strategyn’s approach overlooks this important distinction.

C. *Confused Constructs*

Strategyn further recommends that the same scale be applied to both importance and performance. Yet we’ve observed that using the same scale can cause the respondent to confuse the two constructs, introducing additional error into the data. A better practice is to use separate scales – we typically use a 100-point scale for importance and a 10-point scale, or a letter-grade scale (e.g. A, A-, B+... D-, F) for performance. This clearly distinguishes the two constructs while also providing a more granular view of customer responses.

D. *Missed Opportunities*

Simple modifications to the approach would address these shortcomings, but even so, the opportunity algorithm still suffers from a critical weakness that should give pause to any product developer. The algorithm does perform well in identifying needs with high importance scores and low satisfaction scores. But its structure causes it to break down at the opposite end of the scale, where needs are low in importance. Because the algorithm places a lower limit of zero when subtracting satisfaction from importance, it cannot discriminate between two needs with the same low importance score, but with different scores for performance.

To illustrate consider two relatively unimportant needs: Need A and Need B. Both are of equal (low) importance to customers, and each carries an importance rating of 3. Yet customers are quite satisfied regarding Need A; its top-two boxes score is an “8”. On the other hand, customers are less satisfied regarding Need B, and score it a “3”.

$$\text{Opportunity} = \text{Importance} + \max(\text{Importance} - \text{Satisfaction}, 0)$$

¹⁸ BASES® is a registered trademark of The Nielson Company.

Applying the algorithm to either case yields the same result – the “opportunity” presented by each need is 3. Yet intuitively, the two are not the same. Need A is clearly over-served. Customers are quite pleased with performance, but all things being equal, the need is not important. The right product strategy in this case would be to withhold investment in improving performance, and possibly to engage in a “harvesting” strategy that redirects resources toward more pressing needs.

Need B, on the other hand, also appears unimportant, but customers are not satisfied. In this case, a harvesting strategy may not be appropriate. Moreover, it may be that Need B represents a hidden opportunity to “delight” customers by serving a need they did not believe was important. We have observed on numerous occasions that customers often rate some needs as unimportant *and* poorly performing when they assume that the market simply cannot meet them. But once met, these needs sometimes rise in importance over the course of a year or two. In our experience, many of the most creative new product features begin here.

A Better Approach

For more than 15 years, AMS has used a two-dimensional graphical method (sometimes referred to as a Market Opportunity Map) for displaying importance and performance data (Figure 1) that has far richer interpretability than the Opportunity Algorithm, but without many of the “junk science” characteristics described above. The average scores are calculated for both importance and satisfaction on all of the 15-25 affinized attributes (not the 100+ needs statements that would drown the graph in clutter). Then the “crosshairs” are drawn at the overall mean on each dimension, so that the attributes are likely to be distributed fairly evenly across the four quadrants.

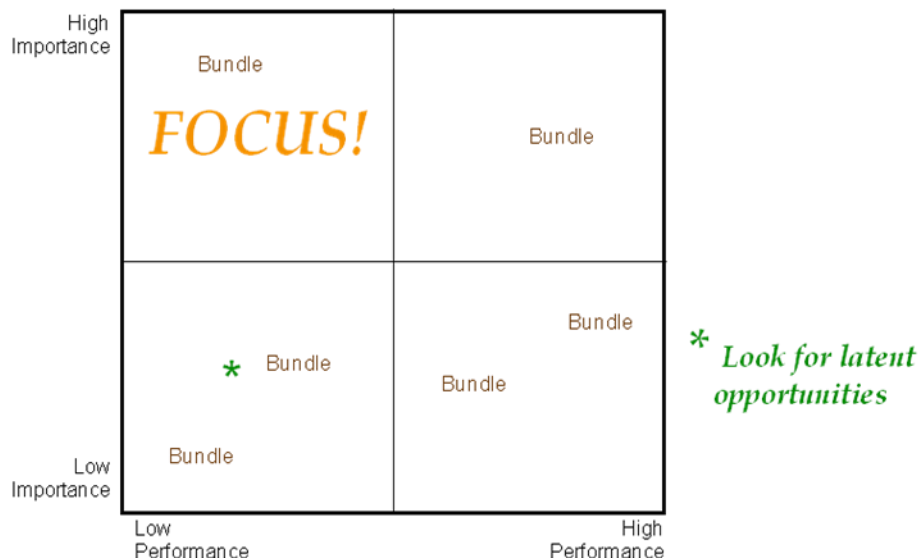


Figure 1: Example of an Importance vs. Performance Grid

An interpretation of the four quadrants is as follows:

Strengths (*High Importance – High Performance*): These are the needs in the upper right-hand corner that customers tell us are very important to them, but they are already relatively satisfied with the options they have available. The appropriate strategy here is to *Maintain* performance at the current levels, at least matching the best of the competition. While improvements here will certainly be valued, they will probably not constitute the primary basis of competition going forward.

Weaknesses (*High Importance – Low Performance*): These are the needs in the upper left-hand corner that customers tell us are very important to them, but they are not currently very satisfied with the options they have available to them. This is the most fertile area in which to *Focus*, i.e. *Invest* and *Innovate*, because fixing these problems will almost always result in significant competitive advantage.

Over-Emphasized (*Low Importance – High Performance*): These are the needs in the lower right-hand corner that customers tell us are not very important to them, and they are already quite satisfied with the options they have available. The most appropriate strategy here may be to *Harvest*, in that there may be opportunities to remove cost without seriously impacting customer satisfaction. Certainly, one should not invest any further in these needs, as improvements here will probably not result in any competitive advantage.

Monitor (*Low Importance – Low Performance*): These are the needs in the lower left-hand corner that customers tell us are not very important to them, and they are currently not very satisfied with the options they have available to them. This is perhaps the most interesting quadrant in that it is sometimes a good place to look for *hidden opportunities*. We have found that respondents sometimes give a low importance rating to such a need simply because they don't believe anyone could ever do better. In these cases, a breakthrough improvement in performance sometimes awakens interest in that need among a sub-segment of the population, and importance suddenly rises in subsequent waves of research. (Using the Kano model analogy, this is the quadrant in which “delighters” are most often found.)

Although the opportunity algorithm is likely to miss these last two quadrants, we have found them to be almost as strategically important as the upper left quadrant, where both methods work quite well. In short, the Importance vs. Performance Grid is both technically superior and more intellectually honest than the Opportunity Algorithm.

Conclusion

In summary, all market research and innovation methodologies have their strengths and weaknesses. In the October, 2001 issue of *Visions*, I published an article entitled “The One Right Way to Gather the Voice of the Customer,” the point of which was that there is no “one right way.” Clearly, some disagree.

Although it wishes to be seen as something radically different, Strategyn is an interesting new entrant into the VOC space. While Strategyn claims to have discovered the “silver bullet” in new product development, most of what they advocate is the same as what we (and many others) have been teaching as “best practices” for years. More troubling, however, is that the technique they propose has a number of disturbing flaws and disadvantages that one should be cautious of in practice. Ulwick argues quite eloquently that his Outcome-Driven Innovation methodology is something new and revolutionary. In truth, it is neither.

+++

ⁱ The author wishes to acknowledge the significant contributions of his colleagues in the preparation of this paper; in particular, John Mitchell and Michelle Harris for their expert help in editing, Steve Gaskin, Jennifer Parr, and Alison Demperio for their skillful execution and analysis of the experiment, and Jennifer Clark and Ben Lodge for their painstaking help with the many references and citations.