

# Expert Panel Discussion on Conjoint Analysis

Sawtooth Software Turbo CBC Event, Monterey, CA

July 2011

*Copyright, Sawtooth Software 2011*

This article presents a polished transcription of a two-hour expert panelist Q&A session that took place at the July 2011 Turbo CBC event, hosted by Sawtooth Software. Bryan Orme from Sawtooth Software served as a moderator/panelist. The other expert panelists were:

- David Bakken (KJT Group)
- Chris Chapman (Microsoft)
- Keith Chrzan (Maritz Research)
- Joel Huber (Duke University)
- Rich Johnson (Sawtooth Software)
- Kevin Karty (Affinova)

**Bryan Orme (Sawtooth Software):** *Let's discuss what we can do to encourage respondents to give more realistic answers that better reflect what they would do in the real world. Joel Huber and Min Ding have presented papers on incentive compatibility. Rich Johnson and I have done papers on adaptive methods of conjoint, looking to engage respondents more and thereby get more realistic answers. There has been some talk about setting the stage for respondents, and putting them in the mood that they would answer more realistically. There have also been some extreme ideas such as paying respondents more if they are more consistent. What in your opinion has worked to encourage respondents to give us not more reliable data in the sense of a fit statistic, but more valid data that is more predictive of what those respondents would do in the real world?*

**Rich Johnson (Sawtooth Software):** I think one of the problems with CBC is that to the respondent it seems repetitive. The respondent thinks he has already answered that question, and why is it being asked again? And one of the big problems is keeping the respondent's interest. We did something in the first version of CBC many years ago, and I think it was a mistake to drop it out, and that was we would start out by setting the stage. Imagine if we were doing a fast-food study, we'd say, "You're driving across the country with your family and everyone is hungry. You come into a city and you think it's time for lunch, and you see these restaurants. Which are you going to choose?" And maybe you'd ask three or four questions like that. Then, we'd say, "Consider this: it's midnight and you've been working hard on

something, the kitchen is empty and you need a snack.” And so on. And I think that kind of thing *did* bring the respondent along and keep him or her more interested.

**Bryan Orme:** It’s not that that is dropped out of the software. You can still do that. Most users think of the “Header 1” of the CBC question as being the fixed thing that shows up every time. Go ahead and leave that blank, and add a “text/html” question on the same page as the random task, and now you can insert whatever prologue you want on top of each CBC question. So it’s still quite possible in the software.

**Rich Johnson:** Yes, but I don’t think people are doing that.

**Bryan Orme:** OK, point taken.

**Joel Huber (Duke University):** Let me add a caveat to that. There was some work done early on related to consuming soft drinks at different times. So, we’re thinking about what kind of soft drink you want early in the morning, and which one do you want later. And, *that* turns out to be much harder than you think it is going to be... that you’re going to have Pepsi in the afternoon, but something else at a different time. People get lost and confused by it. So, you *do* want some sort of scenario that would be kind of a relevant scenario, but it should be just some background noise about some trip that you’re taking which does not lead to one choice or another.

The second thing that I thought Rich was going to talk about—which is a failure, but is a wonderful failure, and it was an important one—was a study that he set up where after a conjoint choice was made, the screen flashed red and made some noise, and it said, “Have you really thought about the question that you just answered? It doesn’t look right.” Or something like that. And then it said, “Now go back and do it right.” And, we thought what we’d find that people would be more consistent afterwards. The computer didn’t interrupt respondents in response to a bad answer. We thought of doing that. But, that’s not the purpose. The purpose was just to say, “This machine is watching you, do you care?” But we couldn’t find any difference in the results. So, that doesn’t work. That would have been great fun had it worked!

**Chris Chapman (Microsoft):** I’d like to make an observation that instead of trying too hard to get CBC right, or to engineer these things to get increasingly precise answers, I’ll often add in other methods, so I’m getting multiple estimates. I’m a strong believer in having multiple indicators.

One observation I have concerns incentive alignment: have people out there actually tried incentive alignment, where you are offering people a real product that they are purchasing out of their gratuity or something? I tried that in one study, and it changes the utilities of the attribute levels. So, it can have a profound effect upon the things you measure. Whether

they're true or not, with or without incentive alignment, I'm not sure. I think the jury is out on that. One thing that I witnessed in practice is that fielding a study like that may run into some extreme difficulties.

In particular, for one study we wanted to see what people would do in a "real purchase" condition. We proposed to give them a larger incentive and asked whether they would spend any of it among certain products. However, the panel and facility provider would not let us do that because it would violate an ethical standard, which is not to do "sales in the guise of research." Instead, they suggested that we mislead people instead – to tell respondents that they were spending money during the session, when in fact they weren't. In the end, the vendor adopted a variant of this, debriefed respondents fully, paid them extra incentive at the end as a surprise, and everyone was happy. In fact the participants asked to do more research like that, because they liked the surprise extra incentive! But this shows how it can be difficult to adapt "real" interventions in a research context.

**David Bakken (KJT Group):** I have a couple of observations. First, a lot of our discussion around making the task more interesting and engaging respondents is based on a presumption that the respondents' cognitions during a choice exercise are something akin to what they actually experience in the real world. However, a lot of evidence on the psychology of survey response suggests that answering a survey question is a unique process in its own right.

My academic background is social psychology. One of the experimental paradigms in social psychology involves setting up a situation where people are deceived to some extent. In this paradigm you don't tell respondents what you are investigating, but you create stimuli that will evoke, based on your hypotheses, some specific effect if the hypothesized process is actually operating. Then, after the fact you tell people, "This is actually what we were interested in." So, social psychologists came under a lot of fire for having people collapse on subway cars to see if people would come to their aid and things like that. Some people felt that this crossed ethical boundaries and I think I probably would agree. But, the point is that the things that people respond to in the real world may not be the things that we are actually capturing in our choice task. And, the more that we understand those cognitive processes and the things that actually do capture respondents' attention, as well as the things that trick them into thinking that they are in the real situation, the more engagement we will get from the respondent.

At the first ART Forum I attended in 1992, Glen Urban introduced Information Acceleration. Whether or not information acceleration is a viable way to get people to think in the future, it does represent an example of attempting to create temporary suspension of disbelief. And, in effect, we're trying to create temporary suspension of disbelief in our choice tasks, asking respondents to, "act as if you are really in this."

Perhaps in part because most of us sitting here are analytically-minded left-brain thinkers, text is fine with us, and we can process grids and things like that easily, so we tend to have very analytically-oriented stimulus materials. Stimulus materials in the real world are much less linear and analytical than what we typically do. And, so I think we are really a long way from having true respondent engagement, meaning that we activate the same processes that are activated in the marketplace.

**Bryan Orme:** As one example that follows from what you just said, I remember a study that a client of ours was doing using our software, and it was on a particular kind of product that keeps people warm at night and they can plug in. They had a conjoint study including all sorts of performance characteristics, because they felt that they had come up with a technically superior product in many ways. And, they did a conjoint which showed that their product really should win in the marketplace against a number of competitors. But, in actuality, the packaging and the way that this product was marketed was nowhere near explaining all the technological benefits of the product as were laid out so clearly in the conjoint grid. So, in reality the product never achieved the market share that the simulator suggested it could. And, so this just shows that when the attributes and the way we lay them out and prime our respondents to think about the superior characteristics of a product—and this often happens when we ask engineers to help us design conjoint studies with attributes and levels—if those are never communicated in the same way in reality, then respondents cannot give us realistic answers in the conjoint task that will align with what they would do in the real world.

**Chris Chapman:** Visualization is something we've often done, especially with product packaging and messaging. One of the nice things is that we can track where people look on the product images. People could click on it and then they could see all the sides of the box, for instance, and drill in. And one of the interesting things that we've observed was a pretty clear definition of two classes of people. There were people who would basically never look at anything, at most just the front side of the box. They might click and see the larger image, and that was it. And, then the second class of people looked at all of the box. I'm not sure if that tells anything generally actionable, but we clearly observed this kind of bimodal behavior in terms of respondent interest in looking at the stimuli.

**Bryan Orme:** *Let's talk about data cleaning. What procedures would we consider best practices for data cleaning? If we clean too much are we throwing out baby with the bathwater? If you do clean, what do you look at? Response time, fit statistics, straightlining? What other things should we look at? What percent of the sample do you typically throw out? Is it 3%, 5%, 20%? Does it depend on whether it is in-person or panel sample? Some researchers use trap questions for non-conjoint data, questions such as "Click on response C*

*for reliability checking purposes.” But, based on the conjoint data alone, how much data cleaning should we do?*

**Keith Chrzan (Maritz Research):** About the “traps,” I just wonder if we do that if we aren’t kidding ourselves about how much they improving data quality. I mean if we annoy our respondents and treat them like idiots, it seems plausible to me that we lose some of their trust and we might get less good data in our conjoint and in the rest of the survey.

**Bryan Orme:** It doesn’t necessarily have to be in a bad way. I see it commonly phrased as, “For consistency purposes to make sure that this is not an automated robot taking this survey, please check box #4.” There are certain ways you can excuse it and say, “I know you’re an intelligent person, but I want to make sure that you’re not an automated robot, so to make sure you’re a human, please answer this.” I don’t think it has to be offensive.

**Chris Chapman:** I still think it is somewhat a waste of time, and I have a personal distaste for it. I’d prefer to design so the problems are minimized. I’ll take an extreme position, that maybe some of the other panelists will respond to, that I don’t think I’ve ever cleaned anyone on the basis of conjoint responses, because I regard that if the design is good, if someone is responding randomly it should have no effect on biasing the outcomes other than to increase the noise. But that’s OK, and I’m willing to live with that. And, if they are answering in a systematically perverse way there is probably no way I can control for that. Now, cleaning people on the basis of non-conjoint items, it’s on a case by case basis. Most frequently, I look at the joint unlikelihood of items that should be unrelated. So, if people are answering several unrelated items in an extremely unlikely fashion, the joint probability of that is quite low. But, in fact few people will achieve these very low joint probabilities; and I might trim these people, such as the top 2% or 5% most unlikely responders, as long as that is based on multiple items and not just one or two unlikely items. But it depends on the goal of the study, and whether I feel confident that we’re primarily interested in central tendency type generalizations

**Joel Huber:** With cheap web data you often don’t know who you are speaking to. And, for those, if you don’t use some kind of tests, you are going to be in deep trouble. On the other side of the equation, being an expensive one, much of the work I do for the EPA uses Knowledge Networks, and these are professional takers of surveys. They are really good at it. We have these awful surveys and in three or four years we’ve had maybe two people out of 5,000 stop. Once they start, they finish, and they finish right. Those two extremes are important because we don’t have to worry at all in one case.

But, you actually have to worry about whether these people are representative of anything in the world. They are professional. They have been called trained seals by some people. They are very good at it. And, they give wonderful data. The r-squares are terrific and everything

works well. But, if you are working with really cheap sample—some are 10 cents a complete—if you want to run a study, it's really quick. And, that's terrific. But, don't expect high quality results. It's pretest stuff.

**Keith Chrzan:** I guess I'd be loathe to clean people out too. The worst result that I usually see is that the utilities get muted a little bit and then my shares get flatter. Unless I had a good reason, unless I could tell that people were being perverse, I won't throw them out. A long time ago I used to go to medical conventions and we would give people card sort conjoints to do. We'd give these people 16 cards to sort. And, these people were doctors and they were doing the card sort so that they could get a Swiss army knife or something. And, in one of our studies, we found that about 5% of our sample not only had terrible fit statistics, but they all had exactly the same pattern of responses. What these doctors had done is that instead of making the tradeoffs we asked them to make, they alphabetized our cards. They went through much more cognitive effort to be smart alecks and to avoid giving us good data. So, in that case I felt perfectly comfortable throwing them out.

**Rich Johnson:** Well I think I disagree about throwing people out. I kind of like to throw people out! A conjoint study involves a model, and what we're doing is taking very elaborate patterns of behavior and trying to reduce it all to something that is described by a few parameters. It isn't always going to work, and when it doesn't work, I don't think we ought to include the individual, because we failed with this person. Rather than put in utilities that are worthless, I think I would rather throw the person out. In my days as a market researcher I was pretty rigorous about that. We would look at the fit statistics for everybody. And, we would usually toss out about the worst 10% of the people. We weren't aiming for 10%—it's just that's about what we ended up with. And, the way we actually did that was to look at a frequency diagram, and to see where the curve bends. And if we could do a good job throwing out fewer, we'd throw out fewer. But we always threw people out that didn't fit the model.

**Chris Chapman:** So, you cut them and then you re-estimated the model?

**Bryan Orme:** There was no HB then, so you just cut.

**Kevin Karty (Affinova):** We throw people out in real-time. They get sent to a window and that's the end of the survey. We use a couple of ways that are really consistent that can be applied to the choice component. If we have four items in a choice task and their placement is random, then the pattern of choices should be random. You can do a simple chi-square test across the tasks to see whether or not it is sufficiently non-random for meeting a threshold to reject that person. Typically we see from 5 to 12% of the people that are voting non-randomly. And, we run multiple metrics and we also see if the person is proceeding at half the median pace. It's important to do the median instead of the mean, so that a couple of people who

stepped off to take care of their kids don't drag on the mean. We see a pretty high correlation between the chi-square test and the median time test. We have a flat-line test too, and all these are pretty highly correlated. The reason we drop respondents is if you include them they tend to flatten the results of the survey—they tend to dilute the differences. Is it a huge effect? It's often not a huge effect, but why do you want any bias? From our perspective, we'd rather get 500 good respondents than 450 good respondents and 50 bad ones, especially since we do not pay our panel providers for cheaters. We also report their userids to the panel provider, and insist that if they do it too many times that they be removed from the panel.

**Joel Huber:** And, panel providers should be doing this—and if they don't they are really hurting themselves. Way back in an early conjoint study we had a summer term, and there weren't many people there, and we got some graduate students to do a study of going to a beach weekend. And, we said to them, we don't care what your preferences are, provided they are consistent. And, we got very complex, very precise R-squares (this was ratings-based conjoint) of 0.96—I mean these guys worked really hard at it, and it *couldn't* have been true. So, you can go too far.

**Bryan Orme:** The fit statistic from HB for CBC studies does not by itself point to a good or bad respondent. A respondent who invents a very simple decision strategy will have extremely good fit, but will have gone through the survey lightning fast. A respondent who was actually thinking a lot harder and realistically, trying to consider lots of attributes, by human nature will make some mistakes and have a lower fit statistic, even though that respondent actually gave you much better data, and is going to work better in your simulator to help predict real-world shares.

**Rich Johnson:** But, on the bottom end, you'd probably agree that a terrible fit statistic doesn't necessarily indicate a good respondent.

**Bryan Orme:** I would agree.

**Chris Chapman:** And one thing I'd add is that I've seen evidence from a bank that did a huge investigation of this, where I happen to know one of the researchers, and the bank went out of business, so they never published this result. The interesting thing was that they went through many different ways that they might identify a bad respondent: internal consistency, failing a trap, speeding, all sorts of things. And, then they implemented all sorts of things for trying to check on that, such as telling the people, "Hey, it looks like you are speeding." The basic conclusion that they arrived at was that between 5 and 15% of people appeared to be responding poorly and failing various of these tests, but that there was no consistent way to determine what to do about that. And, anything they would try would affect the results in some way, but it was unclear whether that effect was a desirable effect on the outcome or not,

and it really was a case by case basis. Looking at speeding may drop some of the people who are responding with a simple but valid decision rule; dropping people who are extremely slow may exclude single moms from the panel or something like that. So, anything will have some sort of effect, you just have to gauge what that is going to be.

**Bryan Orme:** *Back in the day when we used to do in-person card-sort conjoint, everything was done either in-person or mail. Now we have these on-line samples where we can simply pull a trigger, and the data start pouring in. You can collect hundreds of respondents in a matter of minutes because of the economies of scale with these panel providers. Is this a good thing for conjoint analysis?*

**Keith Chrzan:** I like it, and my clients like it; but I'm pretty sure the quality suffers sometimes. When we did things more carefully and took our time, we did a better job.

**Joel Huber:** But, the other side is that you can test and modify your questionnaires quickly. So you do one, and if you aren't sure, you can test it. You can actually do a little hierarchical Bayes, and look at how it's doing. You learn from it. You also learn from giving it to people. But, that ability over an afternoon to get data to test something out was unthinkable before, and it speeds it up. So, think about waves and not just one study.

**David Bakken:** I think there are really two aspects to this question. One is the ease with which we can do more elaborate designs, and if we can have many versions, and we can get better estimates, we can cover more of the design space conveniently. So, there is this great advantage in being able to serve out the surveys via the Internet. On the other hand, I have this feeling (I've been part of a company that sold panel for a while and made that a big part of their business model) that we have tasted this water and it tastes really good and we are deciding to ignore the tainted chemicals in it because it tastes so good. We really don't know a lot about what our samples represent with online panel. We do things to reassure ourselves, but we have a huge non-response problem with online sample and there is a huge number of people out there who don't participate in online research that we just don't know anything about. There are probably only 10 million people that participate in online surveys at all.

**Bryan Orme:** I think in general it's been a good thing, to the degree that companies make sure after drinking the water, as David said, to occasionally get out the chemical set and test it. At the last Sawtooth Software conference, Bob Goodwin from Lifetime Products talked about how they've liked that water a lot, but they've worried about the representativeness of the sample, and also about whether they can good evaluations of their product line without having people sit in the seats and try out the tables, so they tested the traditional mall intercept and compared the results with online samples to make sure the online samples were doing the job.

And based on what he reported at the Sawtooth Software conference, the online samples appeared to be worth it and doing OK for the most part for their product line.

**Chris Chapman:** Related to that, we often do these things in qualitative settings such as during a focus group, and in many cases I've done the same conjoint in a qual setting that I've done online, and then compared the results, and they usually are highly consistent. In terms of the speed of response, I guess I look to panel providers to provide the pure clean water, so having that outsourced is something I rely on. But, in terms of thinking of the speed of obtaining data from pilot studies, it's almost always just a phenomenal thing. I get data back in the same day, and I'm able to look at it. If it is a complete final version of fielding something, I might worry about fast completion of the sample and who is responding, because earlier and later respondents may have different characteristics as to employment or family and the like. But, I usually handle that by asking for a ramp-up period in the sample, and then I also have some levers in terms of quotas and things where I'm in charge in the admin module of loosening up some of that. So, as a practical matter, I try to keep the fire hose shut down a little bit, and then open it up so that things are coming in, and that at least keeps me feeling better. What the real effect of it is, is difficult to assess in general.

**Kevin Karty:** I'm going to say that online panels have been an unabashedly good thing, period, on a lot of dimensions. One of the reasons is we understate the problems we had before we had online sample. We talk about non-response bias, but there was massive non-response bias in mailings. Likewise if you did phone surveys (I come from political science), we know with phone surveys that there are massive social desirability effects. So, "Would you vote for the black candidate?" "Oh, yeah, I'd vote for the black candidate!" But then the black candidate lost. This would consistently happen. It turns out that in online surveys people are more honest because they don't feel the social pressures to conform that they might in a personal focus group or a phone survey, so you have less self-censoring. So, online panels, although they have their problems, fix and improve a lot of the problems you have with off-line methods.

**Bryan Orme:** *What are the most common mistakes being made in the industry in regard to conjoint analysis?* I really liked Joel Huber's comment yesterday when he said the academic hand-wringing that is being done about experimental design probably doesn't matter that much once you have a good randomized experiment with good principles of level balance and near-orthogonality. You can't go much wrong and you can't get much better.

I would say that the area that trips us all up and dictates the success or failure of conjoint analysis studies is the definition of attributes and levels. Making sure the attribute list is realistic enough, finding the right balance of customization: using alternative-specific designs and conditional pricing and maybe a few judicious prohibitions, but especially getting the language of the levels right. Not describing it in language that is too different from what buyers

see in the real world. Not running into the number of levels effect by having one quantitative attribute have 10 levels and another one have 2 levels when you could have easily or somewhat easily have balanced them. Those attribute and level choices will make or break the conjoint study quicker than, for example, experimental design decisions.

**Chris Chapman:** I would add that probably everyone in the room is signed up for some of these online panels, other than the ones that you represent. And, if you are not, do it. I often take surveys from various panels and I see so many incredibly poor designs out there that I give a talk often to undergraduate students in market research classes about them. I bundle these up into the problem of “asking what we want to know rather than what the respondent can tell us.”

In one case, a survey I took was looking at price discounting based on an EPA certification of a home product, and they wanted to know how much more would I pay if this product had an EPA certification versus not. It was a Van Westendorp model—so I guessed that their business problem was that they were trying to figure out whether to pay for this certification or not. As a researcher, it was obvious to me what the business question was they were trying to answer. But, there was no way that I could answer the question in the abstract – “how much more would you pay for EPA certification of a window” because it meant nothing to me. So I said “\$10”, and they told me my answer was invalid. When I tried other values, those were invalid, too. Finally, I guessed at some response that was acceptable to them and let me get through the survey! But that value bore no relation to what they really needed to know from me. And I think this problem is a general problem, especially for those who are new to conjoint, or when things haven’t been pretested. We’re implementing our business question, not necessarily what the respondents can answer.

**Joel Huber:** Let me raise a main-effects versus interactions issue. Researchers come to me and they’ll say, “I have a main-effects design and I want to know about an interaction.” And, sometimes they’re lucky and have four cells (versions) and by combining the cells you can estimate the interactions. But if it is designed to be a main-effects plan—that means the main effects are confounded with interactions—it’s doubly bad. As a general rule, make sure that at least across respondents you can get some estimates of interactions. And, that’s why randomized designs are helpful, because you cannot imagine the kinds of things you are going to want to test, and if it is randomized you are probably fine, and randomized designs cost you very little. So, if somebody comes to you with this neat, clever, absolutely efficient main-effects design, show them the door, it’s not what you want, because that will get you in trouble. And, you’ll never know that you are in trouble because there is no way to test it.

**David Bakken:** I think a category of mistakes arises from giving into clients’ demands to make a choice task look like their view of the product category, perhaps specifying prohibitions that we

should not really do. I will confess to one mistake that I made, in a pricing study where we failed to dig down deep enough to the underlying mechanism of pricing. This study was for high-end office equipment that had a high upfront purchase price, but the customers evaluate these purchases in terms of total cost of ownership: (for example, how much does it cost per page for a copy machine over the lifetime of the machine, taking into account the purchase price, the cost of materials, support, and things like that). I asked the client about total cost of ownership, because I had an inkling that this was an issue. The client said, "We're only interested in the price of the device we are selling... we're not interested in the support cost or anything like that. We've priced our options such that the total price of ownership will be the same across all of our options." So, it's like saying we're going to price these insurance products, but ultimately everything is the same because if you get more benefit over here you have to pay more premium so the cost basis is exactly the same. And, it turned out that there was no effect of price in this model, because there was no difference in the price that mattered to the customers. There was a difference in the price they paid upfront, but the price that mattered was really the total discounted price over the lifetime of the product, and since there was no difference there, there was no reason to buy on the basis of price.

**Rich Johnson:** Well, I would second Bryan Orme's suggestion that the biggest problem is in the front end work that is done with the client to define the attributes and the levels. I think this is very hard to do, because the client usually has a very long list of things he wants to have, and you as the researcher know that the integrity of your study is going to be compromised if you have too many attributes and too many levels. So, there is a negotiation process that has to be gone through, and I think getting that right is the hardest part of the conjoint study.

**Kevin Karty:** I agree, the most challenging part is the definition of attributes and levels, and letting the client force you to create a matrix that has too many attributes, or has attributes that you know don't matter, just because they feel they have to put it in there. Sometimes you can drop an attribute and just ask a regular survey question about it. Another major challenge is the representation of the stimulus. In many situations, we're trying to understand the person's underlying need states and to get deep insight into what's most important and what's not. The layout that we have is a typical grid layout in which all the attributes get an equal amount of attention. In the real world, if your goal is predicting what people are likely to do, that representation doesn't actually match the package on the shelf. Brand name may be huge and the colors are huge, but this particular claim about how the product is made is buried on the back of the label, and so one question is how much we need to make the visual importance of the particular features match the visual importance in a particular ad that consumers would experience in the actual market. And, sometimes it seems we should put a bit more emphasis on the things that are more important in an actual marketing environment.

**Chris Chapman:** I think a lot of the problems we've identified are around the number of levels and pricing concerns and interpreting the None option. One of the things we haven't mentioned much is the problem with MaxDiff. It puts everybody on the same scale assuming that the topic under consideration by the respondents is of the same importance to all the respondents, and it is often not. So, we might rank Swedish Mystery writers where A, B, and C are my preference order. Now my wife might have the same preference order A, B, and C -- but the fact is I read a lot of Swedish mysteries and she reads none. But MaxDiff will happily tell us that we have exactly the same preference structure even though we have quite different absolute preference in the real world. I see people make that mistake an awful lot in terms of interpreting MaxDiff results. Peter Lenk has talked about that a number of times at various conferences.

**Bryan Orme:** That gets into the anchored MaxDiff topic that was presented at the last Sawtooth Software Conference. A lot of people are experimenting with this, and the jury is still out on it. Some people say they love it, and others say it has some problems if you are using the anchor position for each respondent to help define the segments. It's a very interesting area.

***I want to pose another question. Twenty-five and 30 years ago conjoint surveys, with all the lead-up respondent education with attribute glossaries, etc. were often taking about 15 to 30 minutes. Now clients want us to compress the time requirements of the conjoint section to around 5 to 10 minutes. Do the new technologies today including conjoint methods, estimation methods, hierarchical Bayes, and sample sourcing make this OK?***

**Rich Johnson:** The existence of HB makes a whale of a difference in the sense that you don't have to ask so many questions. And, that certainly is a positive contribution. But I don't know of anything that has happened in the last 25 years apart from that that justifies a 5 minute interview where a 30 minute interview was required before. So, it seems to me that if we are only spending 5 minutes in an interview then we aren't getting as much as we used to get with 30 minutes.

**Bryan Orme:** But I can just get six times the sample, because my sample provider can get it to me for pennies on the dollar compared to what we used to pay. So, I just up my sample size and ask fewer questions. Aren't I OK?

**Rich Johnson:** I think you're getting what you paid for.

**Joel Huber:** Conjoint takes about 12 seconds per choice after the first few, so any conjoint that goes beyond six minutes is too long. That to me says it is not the conjoint that is taking the time but what is taking the time are the things that have to come beforehand to get them into the choice. And, there's no way around that—and also getting the subsidiary information. One

of the beauties of many of the panels is they actually have data on people so you don't have to ask that information again. But, getting respondents into the choice is another thing. I don't think it has to last more than 15 minutes. It's 10 to get them into the conjoint, and 5 for the conjoint itself.

**Keith Chrzan:** I'd say the same—we spend a lot of time getting people ready for the conjoint, introducing them to the attributes and levels, showing them a glossary, showing them line drawings, or sometimes a short film even, if it is a new product. Often times we are trying to forecast new products and we have to give our respondents similar experience to what they will face when the product is on the market six months down the road, and the sales rep is trying to sell them on it.

**Chris Chapman:** Often timing is the choice for me between CBC and ACBC. ACBC takes at least a couple of minutes longer, and can be much longer than that. If the question is “can we spend 5 minutes on this or not?” then I'd look at what's the cost in terms of lost opportunity if we do not include this. In general, if I'm faced between having zero choice tasks in the study or having the opportunity to have a few tasks, then I'll take a few. Now, I'll make every argument I can to have the right number, whatever that is for the question, but I would err on the side that something is better than nothing.

**Kevin Karty:** One thing to note about the newer technologies is that it is not just that people are doing fewer choice sets, it's that the technology actually makes the activity faster and easier. So, rather than shuffling pieces of paper it advances to the next screen. If you create dynamic graphics on the interface, you can do things like rollovers and blow things up, and you can communicate things as they are needed. I just wanted to note that the reduction in time is not all about getting less information, or because we are cheating to do things faster, it's just because the administration of the survey is delivered a lot more efficiently, and that is just a plus from everybody's perspective.

**Bryan Orme:** I really like your points, and maybe this will be a little bit of “rah, rah” bragging for the younger generation. We now have the ability to put these surveys on the web with rollovers and grids and pop-ups, and multimedia and laying things out nicely—and I think the younger generation can process that kind of information much more quickly and assimilate that information more effectively than 30 or 40 years ago with the way our minds worked then in terms of looking at and processing information. And, I don't have firm evidence on this opinion, but it strikes me as a distinct possibility.

**Keith Chrzan:** What about the fact that a substantial proportion of respondents are doing our conjoint studies that we are developing to fit on a laptop screen on their hand-held phones? That strikes me as a bit of a problem.

**Joel Huber:** They have good eyes!

**Bryan Orme:** One of the nice things about the SSI Web system is that it saves information about which platform the person completed the survey on, and I think that we are regularly are hearing that anywhere from between 3 to 12 of surveys are being completed on iPhones. And, so you are going to need to be aware of that. People who work with iPhones get used to having to scroll around a little bit to see a full web page. And, maybe they can bring that same kind of proficiency to scrolling and browsing on your bigger choice sets. But, better yet would be if you could sniff out what they are coming at you with and then use appropriate CSS settings to be able to format the discrete choice tasks better for iPhone viewing instead of laptop viewing, and then you adapt to people right on the fly without having to ask them what they are on.

**Chris Chapman:** Has there been research on what happens to responses in these alternate formats? I think about this kind of simultaneous presentation of a choice task that we are accustomed to with these computer screens... there's probably a better way to present that choice task on a smaller device. Has there been any work on that?

**Keith Chrzan:** I was at AAPOR a few weeks back, and there was a whole session devoted to how people respond to surveys on mobile devices. And, it was pretty clear that people don't mind scrolling up and down, but there is a lot of resistance to scrolling sideways. Which, given the layout of a lot of our choice tasks, is a problem.

**Bryan Orme:** *Hierarchical Bayes isn't the only game in town. Let's have a discussion about other models, their strengths and weaknesses. For example, what about probit? We've been hearing a lot about that recently at the ART Forum.*

**Kevin Karty:** Theoretically probit has a lot of advantages. If you look at a normal curve and a logit curve, they look very similar, except logit has slightly fatter tails. The big deal is that a multivariate normal distribution has a covariance. It's like a little hill, and the hill can be stretched out horizontally or stretched out vertically, which says that there is more variance on one axis or the other. It can also be tilted, which means it is correlated. And if it is tilted, it means that if you have high error on one dimension you tend to have a high error on the other dimension. If you have a low error on one dimension then you have a low error on the other dimension. This is different from heterogeneity. Heterogeneity means if there are people that like one thing they also like the other thing. Logit gets that. But what logit doesn't get is that in a particular choice context two items are substitutable for a single person... So, HB probit is adding that one little detail. And, the reason that is potentially important goes back to the IIA

issue, which is that logit can't tell the difference between two items with the same utility that are very similar, and two items with the same utility that are very different. Now, at the upper level HB logit still potentially is differentiating and for a long time people would say, "The upper level pretty much covers it" and you don't need to worry about the individual level error structure. And, some people would say, "Well actually in certain situations you do need to worry about the individual-level error structure." The problem with that is that there is no closed-form solution for it. Now you have a massively complicated model, and it takes a long time to run, and takes a long time to converge. And, it's very nasty in the sense that you don't get a point solution. When you're doing a simulation you have to do draws from your simulation and it's pretty much impossible to code up in a program like Excel, and that's a limitation. A simulation for us is the beginning of a model when we are doing a projection. And, the rest of the model is what is going to happen in market, and what is going to happen with all this other stuff. And all that gets fed through Excel.

So, that's where we net out in terms of how important those theoretical advantages are in reality. I'd ask other people whether they really see an HB probit model deliver substantial advantages over an HB logit model.

**David Bakken:** Well, the HB probit was the alternative to my menu-based choice approach and the computational cost of the HB probit was extreme, while the computational cost of the logit-based approach to menu-based choice is pretty small.

**Bryan Orme:** But that was 10 years ago, right? What about today?

**David Bakken:** Today, I don't know. I don't know what the computational cost of the probit is today.

**Bryan Orme:** What software is available for HB probit commercially? Any?

**David Bakken:** You'd have to work in R I think to build your own.

**Kevin Karty:** R, and it's vastly slower.

**David Bakken:** Depending on your programming skills, you could use C++ or whatever and other things like that. It doesn't have to be R, but there is nothing commercially available that I know of.

**Chris Chapman:** As far as I know anybody using it in production would write their own. So even if they are implementing within an R framework they would write the estimation portion of it in Fortran or C.

**Joel Huber:** Early on I loved probit as a model, and I hated it when I had to use it. It is elegant, it is correct in many ways, if things are multivariate normal, which you'd expect them to be. But, it's a beast to estimate.

I'd like to comment on an issue related to technology and Sawtooth users. The problem Sawtooth users have is anyone can compete with you at very low cost, because Sawtooth is not expensive. And that means that if I'm in your shoes, you want to have something unique. I think Sawtooth's design and data collection is just terrific but you probably want to have some analysis systems which might bring in other kinds of data, might involve elaborate simulators, probit or nested logit, which are relatively easy to work through... but you want something that is hard for somebody to match. If you are not doing that you are going to be competing with other people and it will be hard to distinguish yourself. You're going to want to use Sawtooth, but your shop should be known for something that others cannot do. Maybe such additions represent statistical holy water, but it's really important holy water, and if you don't have any of that, then it's just going to be hard for you to compete. This issue affects academics as well. As an academic you have to have something that makes what you do different.

**Chris Chapman:** I guess I don't disagree in the sense that I think anybody offering services should have some kind of differentiation, so I think that is certainly true. However, responding from the client side, the thing that I always worry about when I see a proprietary method is how do I know it is correct? I know Sawtooth does pretty darn well—and they are so open and transparent about it—but I see something else and I don't know.

**Joel Huber:** I'm with you. If I'm personally a client, then I want plain vanilla conjoint, and it's just great. You wouldn't need anything special. But, some people need more than that.

**David Bakken:** From my perspective, Sawtooth offers the standardization and the transparency that Chris just mentioned, and it frees up resources to do other things. I once worked in an environment which was largely custom. We had a consultant who did the design of the study, then we did estimation, and then we had somebody else who built simulators using Visual Basic, and all this stuff—an entirely custom thing. A lot of resources were put into each of those activities along the way. And, somebody could come along and at very low cost use a tool like CBC 10 years ago, and design a reasonable study on the front end. So, I think your point about differentiation is right. That's why I do all this agent-based simulation.

**Bryan Orme:** *What are the open questions in conjoint research? What are the weaknesses that could be resolved but have not been resolved yet?*

**Joel Huber:** I've mentioned this before, but I'm going to say it again because I want to get somebody to bite on this. In many contexts one choice is what really matters, and when you are setting up a questionnaire, most of your time (10 or 15 minutes) is just setting that first

choice up. So, imagine a conjoint where you have just 12 choices, and the first one rotates. And, so you actually have what would happen if you analyze everyone's first choice. My prediction is that particularly if you make some sort of fuss over the first one and maybe about the last one that you're going to get much better links between those individual choices than to the average partworths. The individual choices will link better to other forms of behavior and to respondent needs and capabilities. The prediction of the preeminence of individual choice is testable; either it will work or it won't. It was confirmed in a heart stent study, where I'm getting much better analysis out of individual questions: the very first one and the very last one, than I'm getting out of looking at the partworths that come out of the conjoint. And, I did not expect that. But, that's what's happening.

**Bryan Orme:** What's so special about the last one that makes it different? Somehow priming them?

**Joel Huber:** In this case, it's only the last one, and you're right, there is a priming effect. In the open heart vs. stent study there are two questions: in the first one we give them an abstract version of two alternatives, and then in the last question we say version A is open heart, version B is stent, now choose, and you get a shift from 80% to 60% going for open heart, just by labeling. Further, if you analyze who does that there is a lot of information in those choices. So, I'm learning more from those individual choices and have a better story to tell.

It's trivial to run a study where you do that, where there is emphasis on the first one and the last one. We historically focus so much on the conjoint partworths, but we need to think about developing skills with individual choices. The big requirement is that individual choices require more subjects, probably 500. But, it's short, it's quick. Anyone who wants to work with me on that I'd be glad to do it. I don't charge anything if I can publish the results.

**Chris Chapman:** This is not really a technical thing with conjoint *per se*, so I'd agree with a lot of things we've talked about here. One thing that has been an increasing ask of me and a frustration that I can't do in conjoint is to include some more emotional or experience-based kinds of things. So, we could think of a certain manufacturer out there who makes lots of incredibly engaging and sleek and fun factors, and those things seem to account for an awful lot of the preference and stickiness of the product line. And, I have wrestled many times with how to implement that sort of design-based experiential or emotional factor into choice models, and I haven't come up with any way in general to think about that or to integrate it very well with other research. But, I see that as a primary limitation in businesses where I'm being asked about experience-based and emotional products.

**Rich Johnson:** I'm not sure that's a conjoint question. I think that's more of an attitude question, it's more of the kind of thing we used to handle under the category of perceptual mapping.

**Chris Chapman:** And that's what I do, but I'm being asked things like, "Well if we make this product more experiential, how will that change our share?" So, I'm being asked a more conjoint-like question for something that is a product attribute in a sense, but it can't be implemented in a conjoint.

**Rich Johnson:** That's very hard, but the APM (Adaptive Perceptual Mapping) that we had for some time claimed to be able to do that. I'm not sure it did it very well, but it made a try.

**David Bakken:** I think part of the problem with things like that is the assumption about how these processes work. The minute somebody comes and asks us to put these things in a questionnaire, there's an implicit assumption that somehow we process these things exactly in the same way that we process the text that describes an attribute, and other things, and that we count all this sort of stuff. In fact, people rapidly process the kinds of stimuli that are involved with design. And what Affinova does with the other part of your business, Kevin, where you are presenting these pairs and you are capturing the real-time optimization by what people choose, is often actually tapping into those processes more effectively than we can do in a conjoint. But, part of it is understanding the elements that trigger our response to design features.

There was a paper presented at ART Forum about using reduced schematic designs for cars and you are basically are trying to come up with the minimal elements of the design that actually trigger our response. So, it's kind of analogous to the way we process faces. We process faces by how far apart the eyes are, the distance from the bridge of the nose to the lips. Our brains are oriented towards these instantaneous judgments based on things like that. And, some designers have figured out the code for some things on those items. But, it is really difficult to translate that into a flat 2-dimensional kind of experience.

**Joel Huber:** This is related to faces. Notice that there is no conjoint on political questions or candidates. What works in those contexts is the thermometer scale. Some sense of how respondents feel about the issues and the candidates—certainly not an additive combination. Again, that moves us in the direction of one evaluation instead of relative evaluations. Kevin knows more about that, but it's a non-conjoint area, but it's an area you're all in, and if you have a client with that kind of need you want to push them in that direction.

**David Bakken:** You remind me, Joel, of the importance of paired comparisons, in terms of our judgments. I'm always drawn back to the vision test that my eye doctor puts me through every two years when I go there. He says, "Is this one better or is this better?" And you are making

the judgment very quickly; it's a comparison of two things. You don't need to articulate the criteria by which it is better, but you do need to be able to perceive something.

And, to the point about politicians, I'm reading a book right now called "why we make mistakes" and it's kind of interesting. One of the book's examples describes research that was done at Princeton. The researchers took black and white photographs of a variety of people and they put them in pairs and asked subjects, "which one do you think is more competent?" The subjects had about 10 seconds to make an answer. The photos were all of congressional candidates. And, the judgment just from the photograph that somebody was more competent was strongly predictive of whether they actually won their congressional race.

**Joel Huber:** Height helps too.

**Kevin Karty:** Regarding using conjoint for emotional issues, there are some things you can do to make it better. One thing is, in creating the stimulus, give some thought to things other than words. So, for instance suppose you are doing brands, and the brand is Mercedes. Well, Mercedes means a lot of things, and one of the things it means is cool and European and sleek. You can create images that convey cool, European, and sleek. One of the things that actually signals quite a bit about brands and products is fonts. Companies spend a lot of time creating new fonts for new product launches. But, there is a library of fonts out there, so you can consider testing things like that. So, you consider ways of testing the overall visual imagery to convey more than just, "This thing is different, but it has the same features as other things." It's more than just features. And, you don't need a very sophisticated tool to do that; you need to think through your stimulus to do that.

**Bryan Orme:** *Realism, virtual reality shopping, graphical representation, actual creating the shopping cart experience on the website, how much do we want to make our conjoint tasks mimic reality, even down to the virtual reality glasses, like you're walking into a virtual reality store, and reaching out for the product, and turning it, any practical experiences about how to increase the realism to somehow get better data?*

**David Bakken:** Yes, I've done some things with virtual reality in virtual environments, and I have to say that I think the biggest challenge is that we have clients that want to have things done in six weeks, eight weeks, four weeks. Creating the virtual reality takes a lot of time. If it's critical enough that the client is willing to allow six months to execute a study, then I think that's great. But, most of the time there doesn't seem to be allowance for time that would enable the development of those materials on a custom basis for most studies.

**Keith Chrzan:** I'm not sure how realistic virtual reality is. When I've done it in tandem with a real shelf test, having people at the physical shelf, I've seen the novelty of the virtual reality is that people will lift up products and they'll spin them and toss them in the air much more often

that they would really do when they are confronted by a shelf, or when we observe people when they are really in retail stores.

**Chris Chapman:** I'd say that these attempts at realism introduce some sort of experimental bias that may be difficult to know what it is, and so I would agree completely with that. One of the things that I often think is an advantage of conjoint is that the minimal presentation strips away effects. In particular the products I've dealt with have always been sold worldwide, and sometimes more outside the US than inside the US. And, retail environments are radically different from one country and the next, and so the question "what is realism?" is hard to know. I think there may be some advantage in actually stripping away some of those things, although as general matter it is hard to say that less realism is actually better. I'd like to say that more realism is better, but then there is the question of what does realism really mean?

**Joel Huber:** We know that if you run an experiment with a real supermarket, the effects are extremely muted. You get just about zero. So, conjoint not actually trying to mimic that. What it mimics is the case of a choice where the information is in front of consumer. That's actually quite different from marketplace choice, and we have to acknowledge that difference. Conjoint makes respondents think about the tradeoffs, but it makes it much more rational than they do in the market. If you want to predict what people buy in supermarkets, it's what they bought last time. That's the answer. That's what most people do.

**Bryan Orme:** Until in the long range, they suddenly say, "Wait a minute! They changed something on me!" and then they change behavior. But it often takes the second or third or fourth time.

**Joel Huber:** Conjoint in some sense simulates that. "I discovered something is different, I'm comparing and am now noticing the differences between them." I think that the virtual reality way of mirroring reality is not the way to go. I'm sad to say this. I would have loved to have found virtual reality working, and I'm glad somebody has tested it. But, it has not done well, as far as I can see.

**Rich Johnson:** I've never done virtual reality, but I think the problem is like this, in the store there are maybe 50 things you could buy. You're interested in one of those, and you want to measure a fairly small change. Well, if there are 50 things that you show a respondent, you're going to get maybe 2% of the choices, and maybe you can increase that from 2% to 4% and you could double your share, but it's too small to measure. So, I think it's much better in an experimental setting to focus on the subset of the products that represent the category for the respondent but which will provide enough choices of your product that you actually have enough sample size to work with. I think there is no way to do this in an attempt to provide a realistic representation of what the respondent would see in the store.

**Bryan Orme:** *Put on your futurist's hat. What's on the horizon for conjoint research? What are the trends and the breakthroughs that are going to propel us into the next decade? I'm interested in what you think is coming, and what is developing that is cool and we're hoping is going to happen over the next 10 years.*

**Kevin Karty:** I think a lot of cool stuff is coming. First of all, I think ACBC is going to increase its share. It's not going to replace CBC completely because there is always a need for plain vanilla. It's cheap, it's fast, and it's mostly right. But, ACBC has some advantages in key areas, such as small panels, more targeted panels, more customization of individual choice sets. And I think that ACBC actually represents a part of a larger trend, which is the integration of different types of data. That one of the things ACBC is doing.

Anchored scales with MaxDiff is also an integration of different types of data. And, I think we are going to see a lot more of that coming, and it's going to be largely successful. There might be some empiricism involved. Academics will probably get upset. But it's going to go there because there will be value there for our clients.

Here are some other big trends, we're going to see integration of software platforms, so we might see things like Adobe Flash developing modules that allow us to custom design interfaces. We're going to see integration of different data sources, and that can include scanner data, it could include some level of biometric feedback, it could include directed responses about particular choice sets. It's going to get to the point where it's almost going to become mandatory, so if you are not participating in that trend people are going to ask why aren't you doing that, it doesn't take that much more effort and so you must be really behind the times if you aren't doing that. And, at some point ACBC is going to be, in many applications, the standard. If competitor X is doing that, how come you guys aren't offering ACBC?

**Chris Chapman:** Something I see is, and this is not conjoint-specific, but I think it is related to this, is the amount of data out there is just enormous. I read that IBM claims there is 10 times as much electronic data today as there was two years ago, and so there is an incredible volume of data from search engines, Google trends, Twitter, etc. There is data everywhere, yet the approach we have in study design and analysis is very often a point in time, cross-sectional kind of view. That view is going to have to change. I don't know how we're going to integrate all these kinds of highly noisy, highly fallible, longitudinal, massive data sets into how we think about things, but it's got to come together. And, I'm sure it will. There is a lot of client pressure to do that, and a lot of opportunity to figure those methods out.

**Bryan Orme:** My answer is certainly going to be biased by the most recent stuff I've been working on. I think that with experimental design, we've gotten just about as good as we can get. I think with utility estimation and market simulation, we're pushing up against a ceiling

and I don't think we're going to go much further that can make a dramatic change for the industry. I think that what's going to be happening is that we'll be needing to customize our questionnaires to be more relevant to the respondent in the sense that they adapt, but also to ask questions in the way that people buy things in the real world. Obviously, that leads to the idea that if things are sold on a menu: bundled vs. a la carte, then why don't we ask that? So, making the questions really mimic what people do in the real world, and then figuring out how to manage the complexity of adaptivity in those designs and managing the complexity of estimating the number of parameters that often expand.

Probably, larger data sets will be more and more common. The nice thing is that panel sample provides the technology over the internet and it is becoming a standard. Certainly mobile technology, and taking surveys on smaller and smaller screens is going to become more of a reality and something we'll need to grapple with, and it is at odds with our desire to have these realistic experiments that need more screen real estate. So, I think that's going to be something that is pushing and pulling us in different directions, and it's going to be a challenge to manage in the future.

**David Bakken:** Building on some of the ideas that have just been expressed, I think that up until now what we do in the conjoint task in the questionnaire has really been estimation-driven. We have this nice framework for how we can figure out these part-worths, so we create a questionnaire that will give us data that fits into that estimation framework. But, as we learn more and more about the cognitive processes that people go through in making judgments it could be that we reorient ourselves to getting the data first, and then using new tools to estimate from the data that make sense, rather than the data fitting our estimation models. And, that could include incorporating a lot of integration of data from different sources as ways essentially to create prior expectations and prior beliefs that we can use to inform our estimation procedures.

**Joel Huber:** I'm going to build on Kevin Karty's brief mention of biometrics. I've done a lot of work on FMRI and I'm very close to a lot of it, and it's not going to help you. I know that. It may if there's a big breakthrough, but there have been so many studies that have shown so little. On the other hand, the low tech versions of biometrics are quite cheap now. You can measure heart rate, sweating, pupil size, and focus of attention. Those are getting much, much cheaper. And, if you need something that makes your shop different from others, that's a really good way to do it. And, it's available now and there has been relatively little work combining the two. I'm going to suspect that within five years they're going to have a computer with a camera that is going to know where you are looking.

**Chris Chapman:** I think they have that now.

**Joel Huber:** Well, not my computer.

**Chris Chapman:** They just haven't told you... (laughter)

**Joel Huber:** I would rather pick the low tech biometric readings rather than the high tech ones, because we know the high tech ones don't work very well. If all you get is heart rate, then you've actually got something. So, there are lots of things you can get, like sweating, that will help measure these emotional things and also tell you how much effort they are putting into it. And, we really don't have the answers to those, so that's a relatively low-cost investment that I would encourage you to do.

**Keith Chrzan:** I believe there's a panel already that has heart rate monitors in some sort of head band that the panelists wear while they fill out your online survey. So, that would be a neat thing to look at.

As far as my own contributions about the future, after these guys who really expect me to come up with something unique, usually when people ask me to forecast, I tell them that if I really knew the future, I wouldn't be at this meeting, I would be on an island and somebody would be feeding me grapes. My feet would be in the water, right? What I do know is what I hear from my clients, and mostly what I hear from them is the rants that their senior managers are asking them, which is why are we doing survey research still? First off, who's answering surveys? What kind of weirdos are answering surveys? Because none of the senior managers who ask this question answer surveys, so they're wondering who answers surveys. There's this massive amount of data already out there in people's own words, why aren't we doing more with that? So, I think that's a challenge that we're going to face as a survey-research industry. At least I'm a survey researcher, that's my background. I worry about the future of the entire survey-research industry, not just conjoint.

**Bryan Orme:** But social media... will that replace the need for conjoint research? Can you really do these complex product tests with multi-attribute problems... can you scrape the internet somehow without producing an experimental design... can you scrape the internet for the words you need to find to be able to answer those questions?

**Keith Chrzan:** Twenty years ago I might have asked, can we really replace high-incidence mail surveys with something that gets one-tenth of one percent response rate? And, the world has told us, "Yeah, sure you can." Will they get the same answer? No. Will they get one that is worth the price they paid? Maybe so.