By Paul Richard McCullough

How not to write a survey in the 21st century

Take a questionnaire written last week and place it side by side with one written 20, 30 years ago. Chances are they will look identical - same logic; same skip patterns; same batteries and scales; same limitations - even though today's questionnaire is most likely being programmed on the Web, with all the new question formats and controls Web surveys offer. Yet the resulting data are often appropriate for nothing more than crosstabs, just like 30 years ago.

Back in the day, quantitative market research meant crosstab decks with 20-point banners. Back in the day, that was rocket science, state-of-the-art, leading-edge. I wrote those surveys (and analyzed their data) with suspender-snapping pride. Problem is, we are no longer back in the day.

Back in the day, corporate mainframes didn't have the computing power of today's smallest laptops. Marketing scientists and other brainiacs have had the last 30 years to develop new analytic techniques to take advantage of all this computing power. These new and not-so-new-anymore methodologies are designed to eliminate many of the biases and inaccuracies of traditional surveys. They deliver answers to questions we didn't even dare ask "back in the day."

But the analytics are just the engine. They need fuel to run. And they need high-octane fuel to run at their optimum. Antiquated survey designs yield very low-octane fuel. They keep these high-powered engines from blowing past the competition and hitting that checkered flag first. Bad survey design turns your Ferrari into a Model T. And it happens every day.

There are three main problem areas in old-school surveys: missing data, collinearity and direct questions. All of these problem areas can be corrected in the survey design, even if you're designing a paper-and-pencil survey, if you understand what types of data modern analytic techniques need.

Missing data

Missing data in survey data sets are epidemic. Don't-knows and skip pat-

Editor's note: Paul Richard McCullough is president of Macro Consulting Inc., a Scotts Valley, Calif., research firm. He can be reached at 831-454-8927 or at richard@macroinc.com. To view this article online, enter article ID 20110302 at quirks.com/articles. This article is an expanded version of an installment of the Beg To Differ column which the author wrote for the spring 2010 issue of Marketing Research *under the title of "Bring Your Survey Design Out of the Dark Ages."*

snapshot

Technology has advanced exponentially in the past three decades but our questionnaire-writing skills have not, the author argues. He cites three problems that must be avoided: missing data, collinearity and direct questions.

terns are the primary culprits here. Generally speaking, both are entirely unnecessary. And both are devastating to advanced analytics.

Many advanced models do not handle missing data very well. Yes, we can attempt to do full-information data imputation and, yes, that is a much better way than mean substitution to address missing data values. But no data imputation technique or any other analytic fudge factor will be as accurate as simply asking everyone the question in the first place. Most questions can be reworded so that skip patterns and don't-knows are not necessary.

The only other alternative is to exclude large segments of your sample because you don't have data for them. This is fine (okay, perhaps tolerable) for crosstabs but when using powerful statistical models to determine big questions - such as "Why do they buy?" - it's important to keep all the sample you can. Not only do you need sample for statistical precision, you want to answer the big questions for everybody, not just for the tiny fraction that accidentally qualified for every skip in the survey.

For example: We've been doing it this way for so long, the logic seems natural:

Q: Do you own any products by Brand X?
If yes, continue
If no, skip next question

Next Q: Please rate this brand on a scale from 1 to 10 where 1 means this statement does not describe Brand X at all and 10 means this statement describes Brand X completely. You may use any number between 1 and 10.

If you feel you don't know enough about Brand X to give it a rating on a particular statement, you can check DON'T KNOW.

Oh, where to begin? Well, let's start with the obvious. Why skip non-owners? We're excluding potentially valuable bits of information by not collecting this data on non-buyers. Oftentimes the client will say they are only interested in how owners rate their brand. But it doesn't really cost any more to skip the skip and ask everyone. Then if you learn something the about non-owners that will help you convert them into owners, who's going to complain?

Occasionally, you may have to change the question wording slightly. So instead of saying, "How would you rate the quality of the Brand X product you own?" you might say, "How would you rate Brand X on quality?"

A slightly less obvious variation on this theme is:

Q: Are you familiar with Brand X?
If yes, continue
If no, skip next question

Next Q: Please rate this brand on a scale from 1 to 10 where 1 means this statement does not describe Brand X at all and 10 means this statement describes Brand X completely. You may use any number between 1 and 10.

If you feel you don't know enough about Brand X to give it a rating on a particular statement, you can check DON'T KNOW.

Now, most researchers will tell you that you can't expect a respondent to rate a brand s/he isn't familiar with. Here's my first problem with that thinking: If you've screened properly so that you are talking to potential and actual buyers of the category, then in the real world, those people will be making purchase decisions about your brand based on the perceptions and beliefs they currently hold, regardless of whether or not they consider themselves familiar with your brand.

In other words, if they are category buyers (or potential buyers), their opinions of you will affect your bottom line, regardless of how well informed they are about your brand. Market research should reflect reality as closely as possible. And poorly- or even incorrectly-informed potential customers are part of reality. Let's measure them. Let's model them. Let's find out why people are (and are not) buying our brand.

My second problem with the above alleged logic is self-assessed familiarity. Some people are insecure. They don't want to commit unless they are certain. With the very best of intentions, they want to provide accurate answers. If they aren't dead sure that Brand X is worthy of an 8 on high-quality, some of them will err on the side of caution and check DON'T KNOW.

Even worse, some people are polite. Faced with the grim prospect of telling some anonymous data analyst that his/her client's brand falls far short on the "all natural ingredients" scale, they rationalize that they haven't eaten Brand X enough to be really sure (they haven't eaten it because they believe it falls far short on the "all natural ingredients" scale) and so they convince themselves the correct answer is DON'T KNOW.

Even respondents who are truly unfamiliar with your brand will have some perceptions and beliefs, even if they have never heard of your brand before. The brand name itself will convey something. These impressions may not even be conscious - they may be registered deep in the subconscious - but they are there. And until they get more familiar, those impressions, however faint, however far above or below the consciousness waterline, will determine whether they buy your brand or not.

All these respondents are making purchase decisions on whatever beliefs and perceptions they do have, whether they're accurate, whether they're based on firsthand experience, whether they're faint whispers in the back of their minds. Let's collect data about reality so we can uncover ways to change it. Note: how to measure subconscious brand perceptions is the subject of another, as yet unwritten, article. It's "beyond the scope" and all that.

Remember, you can always exclude the non-owners or the self-assessed unfamiliar when running crosstabs. Collecting more data doesn't hurt you; not collecting huge chunks of data does.

We're trying to collect data that reflect reality, not a rationalized abstraction of reality. Don't give them the option of saying DON'T KNOW. Make them answer the question!

Collinearity

Any two questions that are highly correlated contain essentially the same information. That is, they are wasting survey real estate. Test virtually any survey data set and you'll find collinearity of epidemic proportions - 100 questions with the information value of 10, if you're lucky.

Item correlation is not inherently evil (like missing values, for exam-

ple; that's always evil). Measurement theory tells us that if we ask a question four different ways and then construct a latent variable based on the four original questions, we will have a more stable, more accurate measure of the underlying theme than any one of the four original questions. So correlation itself is not necessarily bad.

What's bad comes in two flavors:

- Most importantly, correlation that is an artifact of the survey design, rather than inherent statement content, is bad. Really bad, like pushing your little brother down the stairs. You should never do that.
- It's also bad to have those four original questions that are highly correlated and not construct a latent factor. But this is only slightly bad, like putting a whoopee cushion under your little bro's chair at breakfast.

Let's go back to our earlier example. It will illustrate how we often shoot ourselves in the foot writing batteries (or push our brother down the stairs).

Next Q: Please rate this brand on a scale from 1 to 10 where 1 means this statement does not describe Brand X at all and 10 means this statement describes Brand X completely. You may use any number between 1 and 10.

If you feel you don't know enough about Brand X to give it a rating on a particular statement, you can check DON'T KNOW.

TRUST
Is a brand I can trust
Has a good reputation
Is reliable
Been recommended by others

CARING
Cares about me and my needs
Helps me feel safe and secure
Helps me feel confident I've bought what I need
Helps me with guarantees for the "if" in life

PRICE
Offers products that are a good value for the money
Has products that fit my budget
Is not expensive

There are three ways the above battery commits the first (and most important) flavor of bad: 1) adjacency, 2) subtitles and 3) polarity.

Grouping similar items is logical for the survey writer but biasing for the survey taker. By grouping items that appear similar, we're telling the respondent we think they are similar (and they should, too). Correlations will be higher if similar items are adjacent than if they are randomly distributed throughout the battery. A simple solution: Don't place similar items next to each other.

If you take a typical questionnaire and run simple correlations on adjacent items, I'm sure you would find, as I have, a surprising degree of collinearity, even among items that are not similar. The only obvious relationship is often simply their proximity on the page. Adjacency creates collinearity.

Now, I know that subtitles may seem like an obvious no-no to many of you. But I've seen quite a few batteries over the years where the survey writer actually put in subtitles in his/her quest to build sufficient item collinearity to render the battery virtually useless. If adjacency is bad, subtitles are even badder. No subtitles, please.

Polarity is just making all the statements either positive or negative, usually positive. Respondents get in the habit of using a limited part of the scale, typically the higher end (but this varies by culture). By mixing up positive and negative statements, respondents tend to take a little longer to complete the battery because they have to read more carefully, consider each item on its own merits. They have to use a much larger range of the battery scale. Artifact correlations should decrease.

The whoopee cushion flavor of bad (not constructing a latent factor) is bad for a couple reasons: 1) analytic misinformation and 2) inefficiency.

Analytic misinformation can happen a couple ways that I can think of; there may be others. A common practice when determining importance is to take simple pairwise correlations between items and the desired outcome or behavior (e.g., purchase interest). If four items are all highly correlated with each other, their correlations with the desired outcome will likely be similar. All four items may find their way to the top of the list as the most important four items in the survey. The problem is, all four items, because of their mutual correlation, are likely to be measuring the same underlying theme. It's double-counting, or in this example, quadruple-counting.

Interpreting these results can be tricky. If I show four items, all related to product quality, as highly correlated with purchase intent and I show two items related to price equally highly correlated with purchase intent, it is a common and natural error to assume that product quality is more important than price, because there are twice as many quality items as price items in the top 10. In fact, all these data show are that we wrote four items about product quality and we wrote two about price. Analytic misinformation. Not good.

Back in the day, I thought I was hot stuff for building a simple ordinary least squares (OLS) regression model to determine advertising impact on sales. And, in a sense, I was. But there is a danger, particularly today with easy-to-use software, to make an error that leads to an incorrect conclusion. It was true back in the day and it is still true today: Regression models with highly correlated predictor variables are unstable, leading to potentially wildly inaccurate coefficient estimates - so inaccurate that the sign (positive or negative) on a coefficient can actually be reversed. That is, your model can say your coefficient positively drives purchase interest when the exact opposite is true. Analytic misinformation. Still not good.

Inefficiency is easier to explain. If you write four questions that all measure the same thing, more or less, and you don't construct a latent factor that combines the information content of the four questions, then you've essentially spent four times the time and effort collecting one data point than you should have. And that means there were other data points you didn't have time to collect.

If you're going to ask the same question a dozen different ways, don't justify your fuzzy thinking by claiming to be thorough.

Either combine them into a superior variable or admit you're not thorough, you're lazy. Writing good questionnaires is like writing good presentations. It takes more time to write a short one than a long one.

Direct questions

Did you buy that sports car because you want to attract women (Yes/No)? Did you buy my product because of the ad you just saw (Yes/No)? You can bury these types of questions in a check-all-that-apply battery (or whatever else) but you're just putting a dress on a pig. Respondents will answer any question you ask them. But they won't necessarily answer truthfully. Sometimes they don't know. Sometimes they don't want you to know. Advanced analytics can ferret out the truth that respondents may not want or may not be able to share. But you have to ask the questions differently.

The indirect approach is conceptually simple. Ask respondents their attitudes, beliefs and perceptions. Ask them some measure of the desired behavior. That might be recent past behaviors such as purchase, visiting a Web site, making a donation. It could be a claimed likelihood measure such as purchase intent. In general, the more concrete the better. Actual behavior is always going to be more useful than claimed behavior. But we don't always have actual behavior data available.

Either way, indirectly deriving importance involves modeling respondent characteristics such as attitudes, beliefs and perceptions as predictor variables with some desired outcome, such as product purchase, as the dependent. There are a variety of ways to attempt this but in its simplest form, at least for the purposes of illustration, think of an OLS regression model. That will give you the idea. In practice it can get a little more complicated.

But the outcome is always the same: those respondent characteristics such as his/her attitudes, beliefs and perceptions that best explain the variance in the dependent variable are more important than those that do not.

Ask a male respondent how important the Playboy channel is to his decision to buy the premium package from his cable company and you're likely to get very low importance scores. This was even more true when we did mall interviews with college coeds as interviewers.

But conduct a choice-based conjoint analysis and you might find a different answer entirely. Why? Choice-based conjoint derives the importance of the Playboy channel by analyzing the pattern of responses across a wide range of programming options. It's indirect. The respondent isn't aware (and neither is that coed administering the interview) that his answers will ultimately reveal his true motivations.

When it comes to advanced analytics, direct questions have another, albeit less common, downside. As predictor variables in a model, they're useless. Typically, advanced analytics involves modeling the data set to determine what drives some behavior. There are lots of other questions to ask, but this is the big one. Asking respondents how important certain features are to their purchase decision is a direct way to get at the same answers the model is trying to uncover indirectly. The problem is it is very difficult to put importance data into a causal model and make any sense of it. Suppose I put brand imagery ratings in a model and I conclude that the higher a respondent rates Car Brand X on crash safety, the likelier the respondent is to buy the car. In other words, perceptions of Car Brand X crash safety drives purchase intent. But what if I didn't rate Car Brand X on crash safety but I rated the importance of crash safety in general? Even if I believed the data (which I wouldn't - this guy wants to attract women), how do I interpret that? The more importance a respondent places on crash safety, the likelier he is to buy the car? Really? Even if he thinks the car is flimsy as a cardboard box?

Why would anyone want to cram the square peg that is stated importance data in the round hole of a causal model, you ask? I'm not really sure. But I have been asked to do so on numerous occasions.

I think the process goes something like this: a researcher is awarded a project and writes a questionnaire the same way s/he always does; s/he copies and pastes from the last study. Importance batteries are standard fare. Then after the fact, just about the time rigor mortis is beginning to take over the data set, someone says, typically in desperation, "We haven't got a story yet. Let's build a driver analysis model." And what data do we have to put in said model? Yeah, stated importance. And, of course, running a model with no theoretical justification just about always gives you some spurious correlations to scratch your head over.

Miscellaneous other

I haven't yet addressed monadic scales. They don't fit neatly into my three problem categories of missing values, collinearity and direct questions. But they are a mainstay of questionnaire design and they have to go.

There is sufficient high-quality literature on the problems with monadic scales to make the debate officially over. Monadic scales are almost useless. There are typically three main issues that must be addressed: minimal variance across items, i.e., flat responses (huge problem); brand halo (largely ignored, but that doesn't make it go away); scale usage bias (also ignored).

Resulting data are typically non-discriminating, highly correlated and potentially misleading. With high collinearity, derived importance scores may actually have reversed signs, leading to absurd conclusions (e.g., lower quality increases purchase interest [see collinearity section above]).

The solution is to avoid monadic scales entirely if at all possible. Max-diff is probably the best alternative in most situations. There are some limitations with max-diff that currently make it difficult to apply to brand imagery measurement but there is work currently being done in that area. Without getting into the gritty details, if you want to apply max-diff to multiple items, like several brands, you could look into dual-response max-diff, the latest innovation in max-diff scaling, or some data fusion techniques. Both hold some promise here.

If your scaling needs involve

just one item, such as an importance battery, max-diff is definitely the way to go.

Frustration has been growing

Although my frustration at being asked (repeatedly) to administer CPR to data sets postmortem has been growing for many years, this article was inspired by just one recent questionnaire. It was not different from but representative of generally well-regarded survey design. It was a typical survey written by smart, experienced researchers.

I'm sure that I have only discussed the tip of the iceberg and that there are numerous other egregious errors that need to be identified and removed from modern-day questionnaire design that I haven't mentioned or yet discovered. If I reviewed a dozen past surveys I'm sure I'd have a longer article.

I bet you can think of other questions you've run across that create biased or misleading results simply because of the way the question was written.

For example, one problem question that I discovered in my muse survey didn't fit any of the three categories I listed above. It is a very common question type, too. It was a "check three" question. In this case, it was an importance question, i.e., "Check the three most important attributes when deciding to…"

Imagine this scenario: For simplicity, half our sample all feels the same way (no heterogeneity within that half). And how they feel is there are four important attributes that influence their decision to do whatever it is the client wanted them to do. One attribute (the same one, Attribute D) always gets left out in the "check three" question. This half all makes the desired decision (e.g., they bought the product, subscribed to the service, called the 800-number, visited the Web site, etc.). The other half picks all of the attributes with equal likelihood and never makes the desired decision.

Let's look at the correlations. At least half the respondents who checked Attribute A made the decision the client wanted. Almost all the respondents who did not check Attribute A did not. Same for Attributes B and C. High degree of positive correlation between Attributes A, B and C with the desired decision. What about Attribute D? All respondents who checked Attribute D did not make the desired decision. At least half the respondents who did not check Attribute D did make the desired decision. High degree of negative correlation, even though Attribute D is, in fact, highly correlated with the desired decision. By limiting the number of attributes to be checked, we created the opportunity for a spurious negative correlation. I saw this negative correlation in a real data set.

Solution? Well, by now you know how I feel about direct importance questions and monadic scales. It is preferable, in my opinion, to collect the appropriate data and build a causal model, deriving importance based on the correlations between attitudes, beliefs, perceptions and the desired behavior. But if you must, use max-diff. Don't use "check three."

Understand how the data will be used

Modern marketing science offers us the chance to see a little more clearly, dig a little deeper, forecast a little more accurately. In some cases, it's not a little. It's a lot. We have to understand, however, how the data will be used prior to writing the questionnaire so we can collect data appropriate for the subsequent analysis.

Even without fully understanding the analytic plan, following these simple guidelines will vastly improve the quality of your data and subsequent analysis: avoid missing values by eliminating skip patterns and don't-knows; prevent collinearity by mixing things up (item order, polarity, etc.); derive importances - don't ask directly; and avoid monadic scales whenever possible (it's not always possible just yet). |Q