.50
+4.25
+1.25
+.50
+7.50
+2.
–.25
–1.50
–.75
– 2.00
–1.25

The Hidden Limitations of Tracking Research

By John Seal and Mark Moody

Stop wasting money by tracking the wrong measures.

Successful organizations place a high priority on monitoring the impact of their marketing efforts. They seek evidence of the impact of their actions when they are successful and hope to identify downturns in performance or perception in a timely fashion so that negative business results can be averted. Consequently, a large portion of many marketing research budgets is spent on tracking studies.

Tracking studies are often the most visible products of research departments throughout the organization and are critical for tracking customer loyalty as a leading indicator of defections. They can also track brand and advertising awareness to measure the effectiveness of advertising campaigns. Finally, tracking studies help track brand image to monitor investments made in brand-building efforts and to understand evolving consumer perceptions of brands and products.

Inside Research (May 2006) reports that of the $6.8 billion spent on marketing research annually in the United States, 5.5 percent, or about $375 million, was spent on brand and advertising tracking studies. The March 2007 issue further reports that more than $650 million is spent annually on customer satisfaction measurement in the U.S., and much that is tracking research as well. Unfortunately, much of the investment devoted to generating these critical streams of information is not utilized as efficiently as it could be.

Executive Summary

Tracking research, including brand/advertising awareness and customer loyalty tracking, represents substantial investment. Tracking projects can involve many statistical comparisons that are not actually meaningful. The power and reliability of these projects can often be simply and inexpensively increased by tracking means rather than "top-two-box" and other similar percentages. Reporting more sensitive metrics, rather than familiar but less quantitatively rich measures, can help to overcome hidden limitations of tracking research.

Tracking studies typical involve measuring a large number of survey variables over several time periods. Often the critical information is not the absolute level of the measures, but how they may have changed as compared to a previous time period. Consider the following hypothetical, but realistic, tracking program design:

- A total of 1,000 interviews are completed each quarter.
- The sample is allocated equally over 10 markets (100 per market).
- Reports are generated each quarter comparing period-to-period results on 40 attributes in total, by market, and by each of three target segments for our brand and our major competitor.

Under this scenario, each quarter's reports could involve 1,120 comparisons to the prior quarter:

(10 markets + 3 segments + 1 total) * 40 attributes * 2 brands

In situations like this, analysts typically employ statistical significance testing to help make sense of the large volume of results and better understand which changes are "real," meaningful, and worthy of management action vs. simply random variation that represents no real change.

Survey-Wide Reliability

Statistical tests are familiar tools to researchers, but users of information often forget what a finding of "statistically significant" really indicates. Researchers fall into the habit of treating statistically significant as equivalent to meaningful. Recall what a "90 percent confidence level" on a series of statistical tests does and does not mean.

It does not mean that 90 percent of the significant findings are large enough to warrant action. (The magnitude of difference required to be "significant" is a direct function of the sample size selected for a study and has little bearing on managerial meaningfulness.)

Further, it does not mean that 90 percent of significant findings are "real" findings, as opposed to being due to chance variation in sampling. (This is a common misconception, but, as we shall see, the true percentage of "real" differences can be far smaller.)

What a 90 percent confidence level really does mean is that, if there were no actual changes at all, a correct conclusion of "not statistically significant" would typically be reached for 90 percent of the tests conducted (and thus the remaining 10 percent would be identified spuriously as significant changes, merely due to random variation in the types of individuals who were included in the research sample in different quarters).

Consider the implications for the hypothetical tracking study introduced earlier. If 1,120 period-to-period statistical tests are conducted using a 90 percent confidence level and nothing really changed at all, approximately 10 percent, or 112, of the comparisons would be expected to be statistically significant.

But typically some measures undergo real change between measurement periods. Suppose this project was executed as described and 142 significant period-to-period differences were detected. That finding would suggest there were probably 30 "real" significant differences along with the baseline expectation of 112 that did not represent an actual change, but were just a consequence of random differences between samples.

Thus only 21 percent [30/142] of the changes reported as significant would likely have been "real," far lower than the 90 percent that many researchers erroneously believe a "90 percent confidence level" implies. It is of course impossible to know which 21 percent are the real changes, so 79 percent of the conclusions researchers draw, and actions managers take as a result of those conclusions, may simply be wrong.

The 21 percent figure in this example is referred to as "survey-wide reliability." It indicates the percentage of research conclusions that are reliable. Survey-wide reliability is a powerful concept and a measure that too often no one bothers to calculate. When it is calculated, it can illuminate some unpleasant truths about the true amount of reliable information produced by expensive tracking studies.

Survey-wide reliability is related to the notion of experiment-wise error rate. When a series of statistical tests are conducted together, each one individually has an error probability set by the confidence level of the test (a 90 percent test will have a 10 percent probability of a Type I error). However, the 10 percent probability of error in each individual test compounds, so across the entire set of tests the probability of an erroneous conclusion on at least one test can be far greater than 10 percent. In fact, in a large tracker like the one described here with upwards of 1,000 tests, the probability of at least one Type I error approaches 100 percent.

Statisticians offer a number of techniques to adjust tests to account for experiment-wise error rates, but they are rarely used in marketing research. While very useful in making one experiment that involved a single test comparable to another experiment that involved a handful of tests, these adjustments are really not practical or useful when hundreds of tests are being conducted simultaneously.

Survey-wide reliability can be calculated for any tracking study simply by adding up the total number of tests conducted, totaling the number of significant differences observed, and determining the expected false-alarm level. A review of a handful of recent tracking projects yielded an average survey-wide reliability of just 9 percent!

How can the results be so bad? These projects were drawn from product and service categories in which very little real change was occurring from period to period, and each of them reported results quarterly or more frequently, allowing very little time for real change to occur. The low reliability figure is likely typical of programs with these characteristics. However, when no real changes are occurring, researchers have a tendency to fall back on the false alarms to give them something to talk about.

Researchers have proven themselves very capable of coming up with insightful explanations for nearly anything, and they expend a non-trivial amount of time generating "stories" around spurious results. Better yet, whatever mitigating response management makes to surprising downturns in their scores will generally be rewarded in the very next measurement period, as random fluctuation in period-to-period results leads to a "correction" back to the true level.

Beyond providing informative assessment of the usefulness of tracking studies, survey-wide reliability can also serve as a valuable metric to evaluate different strategies in tracking study design. A design that leads to greater survey-wide reliability, especially at minimal additional cost, is likely a design that will bring more useful "bang for the buck."

Better Strategies for Tracking Studies

Researchers tend to think of statistical tests, and their conclusions regarding statistical significance, as a referendum on the size of the difference they are testing: The bigger the difference, the more likely it is to be statistically significant. And that is true—if all else is equal. But all else is not necessarily equal; other variables in the equation can be manipulated to derive more value from tracking studies.

Ignoring minor differences between different statistical testing formulas appropriate for different kinds of data, all significance test formulas are basically the same and have the same three basic elements:

$$\frac{\text{Size of difference}}{\text{Variability between individuals/Sample size}}$$

Because a larger value computed according to this formula leads to a greater chance of a difference being deemed significant (for a given confidence level), more significant findings will be identified when (1) the size of the difference is larger, (2) the variability between individuals is smaller (relative to the size of the difference), or (3) the sample size is larger.

Researchers cannot much influence the real size of the difference through how they design their research programs. But while it's easy to think of the statistical test as a function of the size of the difference, that's really just one element of the formula.

Researchers do have the ability to manipulate the other two elements to some degree when designing and planning tracking studies in order to increase survey-wide reliability and get the most value for their research dollars.

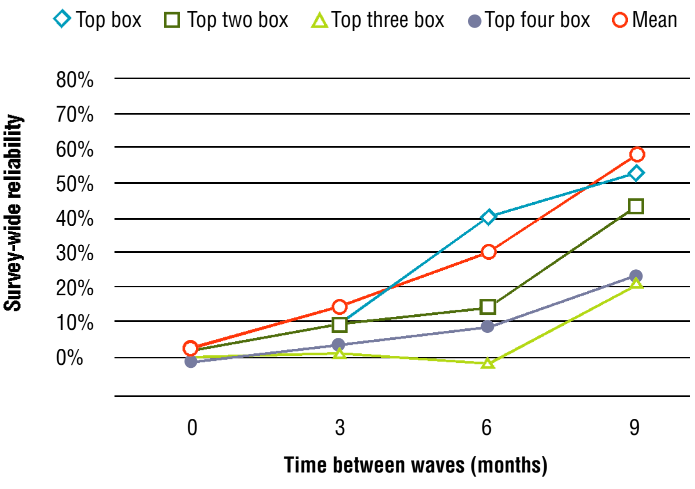
Variability between Individuals

In tracking studies, researchers typically prefer to report percentages. When tracking a measure like brand awareness, a percentage is a very proper and logical metric to use. But even when they have measured a variable in a richer fashion, such as a specific performance attribute on a 5- or 10-point scale, standard practice seems to be to nevertheless "dichotomize" the multipoint scale into a percentage, such as the percentage of respondents giving a "top-box" or "top-two-box" response. Reporting the "top-two-box" effectively turns a 5- or 10-point scale into a two-point scale; each individual respondent either gives a top-two-box response ("1") or does not ("0").

Instead, researchers should consider the mean as the primary measure to track on multipoint scale attribute ratings. The reason usually given for dichotomizing rating scales into a top-two-box percentage is that percentages are more meaningful to management. Saying 65 percent of individuals responded in a certain way is more concrete and easier to grasp and react to than reporting than the average rating was 3.6. There is some merit to this argument, but at what cost?

The mean is not a terribly complicated or unfamiliar mathematical concept for professional decision makers to understand, and it has much to support its use, even beyond the topic of this article. For instance, researchers often use tracking (or other) data to build predictive models in which the relationship between individual attribute ratings and some overall measure (like loyalty or brand affinity) is assessed.

Exhibit 1 Estimates of survey-wide reliability with different sample sizes



Based on Monte Carlo simulations of n=500

These models typically use regression methods, and the parameter estimates (betas) they produce are generally interpreted in terms of the relationship between measures.

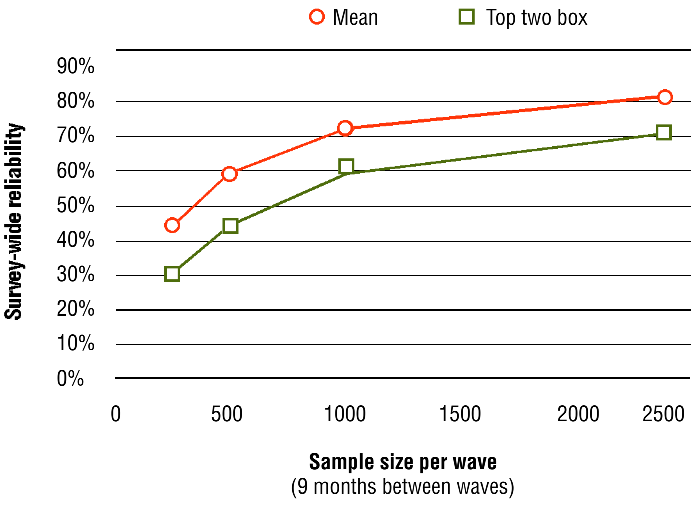
For instance, a one-point change in an attribute may be associated with a certain size of change in the overall measure. When analysts make interpretations like this, they are referring implicitly to one-point changes in attribute means. If predictive models are going to be used to provide useful information about the impact of changes in means, these are the metrics that management should be most familiar and comfortable with.

But more important for this discussion, dichotomizing data from a 10-point scale to what is really a two-point scale effectively increases the variance, or variability between individuals. Quite simply, if every individual is either in the top-two-box or is not, then each individual is at one or the other of the opposite ends of the two-point scale. Two individuals who have different scores are as different as they can possibly be, on the scale used; there is no middle ground.

In contrast, retaining a full 10-point scale (for example) allows for many shades of difference in strength of perception. Two individuals can differ by only a small amount, not necessarily by the full length of the scale, yielding less variance (relative to the length of the scale and the size of differences measured). Thus calculating a mean that incorporates information from the entire scale has the opposite effect from dichotomizing into a top-two-box summary measure; it decreases the variability between individuals, which, as seen earlier, acts to provide more powerful statistical tests (at no real financial cost) and increase the survey-wide reliability.

Improvement in survey-wide reliability can be illustrated through Monte Carlo simulations on real data. Monte Carlo (or "bootstrapping") simulation methods, as applied here, involve drawing repeated samples from a population determined by real data in order to simulate running the same experiment many times. Doing so provides stable estimates of measures like survey-wide reliability that do not lend themselves to simple statistical formulations.

Exhibit 2 Simulations using sample sizes other than 500



Based on Monte Carlo simulations

In one quarterly tracking project conducted by Burke Inc., 20 attributes were rated on a five-point agreement scale. Nine of the 20 showed statistically significant differences (at a 90 percent confidence level) between the first and fourth quarters. Generally, these attributes were closely associated with the specific investments made by the firm during the year. Thus, there were real effects to be found.

Monte Carlo sampling was used to calculate the survey-wide reliability across the following simulation parameters:

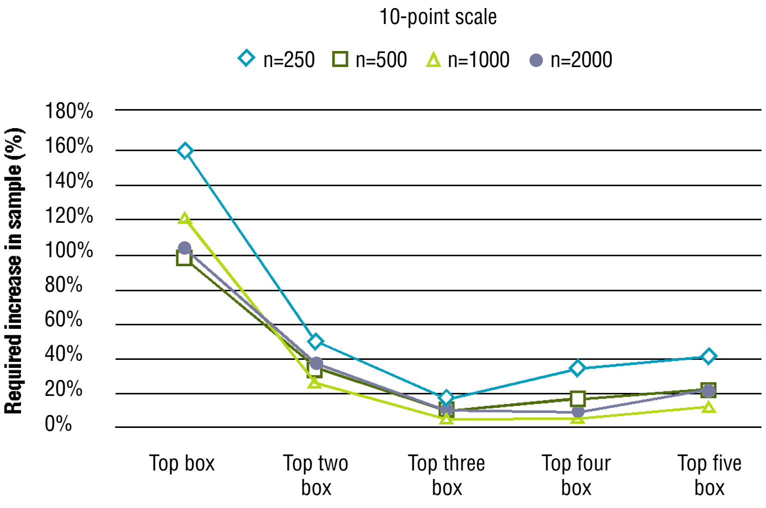
- Time between waves of zero, three, six or nine months. (The zero time represents a control group simulation in which samples are randomly assigned to one of two groups and then tested for a difference. As validation of the Monte Carlo approach, the zero time analysis should produce merely the expected false-alarm levels, or zero percent survey-wide reliability.)
- Measures: top-box, top-two-box, top-three-box, top-four-box, and mean. On average these metrics were: top-box=20 percent, top-two-box=61 percent, top-three-box=90 percent, top-four-box=98 percent, mean=3.7.

Each of the 20 combinations of these two parameters was simulated 3,000 times, each time drawing a different random sample of 500 individuals (with replacement, from a full sample of approximately 2,375 per wave), in order to generate stable estimates of survey-wide reliability with different sample sizes. Exhibit 1 illustrates this.

The best-performing measures are the mean and the top-box, both performing substantially better (yielding more significant differences with the same expected baseline number of "erroneous" significant differences) than the traditional top-two-box. Further, the size of their advantage over other measures grows over time. The poor performance of the top-two-box may be due to its 61 percent average score. If 61 percent of consumers are already giving a product or company the best possible score (after dichotomizing the measure), then the ability to measure gains is largely restricted to the remaining 39 percent. In contrast, with only a 20 percent top-box score, improvement could be measured among 80 percent of the sample.

High absolute scores on key measurement metrics are often seen as desirable. Managers may prefer to see themselves achieving well on key measures, as reflected by a "high" percentage value, thus the popularity of the top-two-box measure. However, this approach dooms the user to detecting less of the change that may be occurring, albeit while congratulating themselves on having such high baseline scores. (In extreme cases, the mean could also suffer from a "ceiling effect" with little room for improvement if a large portion of the sample gave responses near the top end of the scale. However, in cases like that, top-box and top-two-box measures would be similarly or even more severely affected.)

Exhibit 3 Increased sample required for metrics other than mean



Percent increase required to achieve same survey-wide reliability as mean

Impact on the Bottom Line

The remaining element to address in the formulas used to conduct statistical tests is the sample size. Simply employing a larger sample size is a more familiar and intuitive way to produce more sensitive statistical tests and generally higher survey-wide reliability (because more "real" statistical differences are likely to be detected, given the same baseline number of expected "random" differences).

Consequently, it is possible to assess the real gain from using an efficient metric like the mean rather than the standard top-two-box percentage by using the common currency of survey-wide reliability. Monte Carlo simulations can be employed once again to compute the equivalent sample size gains of using the mean over alternative metrics, illustrating the practical benefit of optimal metric choice in real terms.

Exhibit 2 shows the result of Monte Carlo simulations using sample sizes other than the 500 used in Exhibit 1. The findings suggest that it would be necessary to double the sample size to achieve the same survey-wide reliability when tracking top-two-box as simply tracking the mean would provide. To illustrate, the survey-wide reliability using a mean measured on a sample of 500 is 59 percent. This same 59 percent reliability would be attained only with a sample of 1,000 when tracking top-two-box percentage.

These results demonstrate that, for at least this one program, tracking the mean rather than top-two-box is equivalent to doubling the sample size at no extra cost—a potentially startling conclusion.

Of course, this conclusion is based on just one data set, and not all studies use a five-point scale for attribute ratings. To attempt to further generalize the findings, similar Monte Carlo simulations were performed based on a different data set, in which attributes were rated on a 10-point scale.

The findings of this analysis suggest that it is not universal that the top-box score performs similarly to the mean as seen earlier; in this case, top-box performed more poorly than other "top-k-box" measures, but the mean continued to perform the best in terms of survey-wide reliability.

Exhibit 3 shows the percentage by which sample size would have to have been increased for the various top-k-box measures in order to achieve the survey-wide reliability that would be achieved by merely using the mean as the primary reporting metric.

While top-three-box comes closest to the mean in survey-wide reliability, it still entails a sample size penalty of 6 percent to 16 percent, depending on the sample size of the test involved. The penalty for using the top-two-box is substantial, as sample sizes would have to be increased by 25 percent to 50 percent to stay on par with the mean.

Using the conservative end of this range (25 percent), and adding the further assumption that half of tracking study expenditures are variable costs related to sample size, a quick calculation reveals that 12 percent (50 percent of 25 percent) of tracking study spending could be wasted merely because researchers are choosing to report the top-two-box percentage instead of the mean.

Based on the *Inside Research* estimates cited earlier, as much as $800 million is spent annually on tracking research in the United States alone. If all of that were spent on programs tracking the top-two-box percentage rather than the mean, the 12 percent translates to about $100 million wasted due to simply choosing the wrong metric to report. Reallocating these monies to larger sample sizes or to additional research initiatives could increase the visibility and contributions of research functions in many organizations.

The inefficiency introduced by choosing less-than-optimal reporting metrics costs researchers money in real terms, because the only path to recoup the lost statistical power is to spend more money to conduct more interviews. These analyses demonstrate that sample size and optimal metrics are two paths to achieve the same goals.

Researchers need to be aware of the surprisingly low percentage of statistically significant "findings" in tracking studies that likely represent true meaningful changes in the marketplace. Reporting more sensitive metrics, such as the mean rather than top-two-box and other familiar but less quantitatively rich measures, can help to overcome this hidden limitation of many tracking programs. ●

John Seal is vice president and senior consultant in the Decision Sciences group at Burke, Inc. in Cincinnati, Ohio. He may be reached at john.seal@burke.com. **Mark Moody**, PhD, is a retired Vice President and Senior Consultant at Burke, Incorporated. He currently serves as consultant to Burke on an ad hoc basis, and he can be reached at mark_moody@bellsouth.net.k