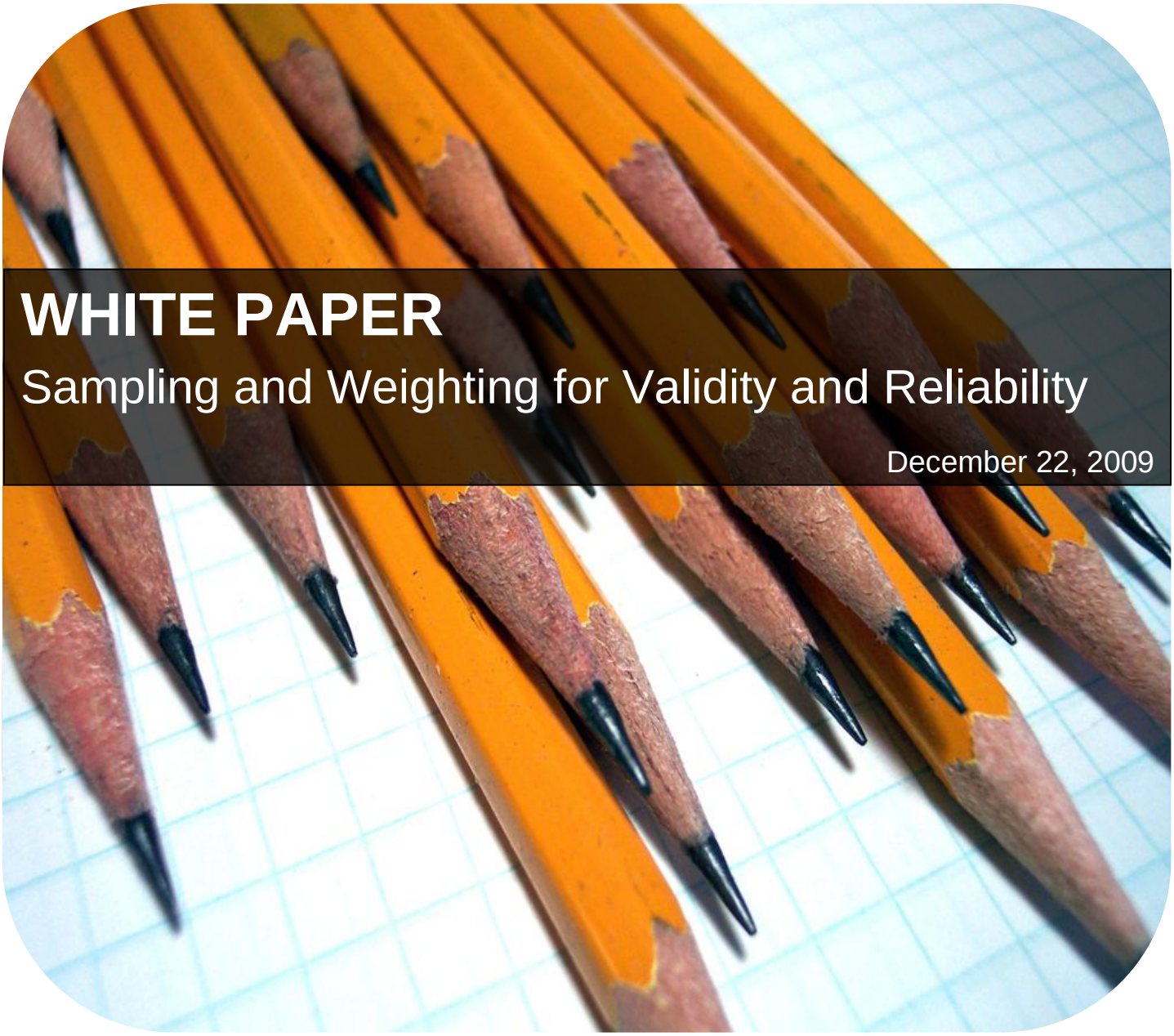
conversition™



WHITE PAPER

Sampling and Weighting for Validity and Reliability

December 22, 2009

What is Sampling?

Sampling is a technique used in research for selecting people (or things) to participate in or provide data for research purposes. It is a vital component of any research study that can directly affect the validity and reliability of research results.

A famous example of poor sampling that continues to be discussed today occurred in 1936 when the *Literary Digest* predicted that Landon would win the US election with 55% of the vote. In fact, Landon lost after getting just 37% of the vote. Even though this survey was based on a sample of 10 million people, the results were completely wrong. Why? Because the sampling method relied on a telephone book and vehicle registration records to select the survey participants. In 1936, this meant that only wealthy Americans, who were more likely to be fans of Landon, could participant in the survey. (By the way, Roosevelt won with 61% of the votes).

Sampling is almost always used in research because it is neither financially nor logistically feasible to survey every single individual in a research population. In the case of social media research, it is similarly impossible to find and collect every single verbatim on the internet. Effective sampling ensures that results from a small portion of the available data are just as valid and reliable as results from the entire population. As such, a well developed sampling plan ensures that the full range of opinions and perspectives has a fair opportunity to be included in the research.

Traditionally, sampling is used to identify which groups of people, perhaps business owners or women who have teenagers, will provide the data. In the social media research space, sampling doesn't refer to the selection of individuals, but rather to the selection of website sources. As such, the sampling frame for social media research consists of all public blog entries, status updates, product evaluations, video comments, and written contributions on the internet.

Identifying people to participate in research is easy. Identifying the *right* people is sampling.

What is Weighting?

Weighting is another procedure commonly used in marketing research to promote the reliability and validity of results. Weighting means that results from certain subgroups of people are given more or less prominence when the results are being calculated. For example, if the intention is to produce results that reflect the general population, it is important that the sample consists of men and women in equal proportions. But, if that didn't actually happen, the results can be statistically weighted to simulate a dataset that did in fact include equal numbers of men and women.

In the case of social media data, weighting ensures that when the quantity of data coming from different sources changes, that change won't bias the results. This is particularly relevant when websites make new or additional pieces of data publicly available. Just because a source now makes twice as much data available does not mean that its data should count twice as much.

Let's consider a theoretical example. Consider a source of data that creates 1 million records every day but only releases 10 000 of them every day. After numerous months, though, the website starts releasing 100 000 records every day. The website has not gained any more users. The users have not been creating more records. The opinions of the users have not changed. The only change is that the source has increased the access to records.

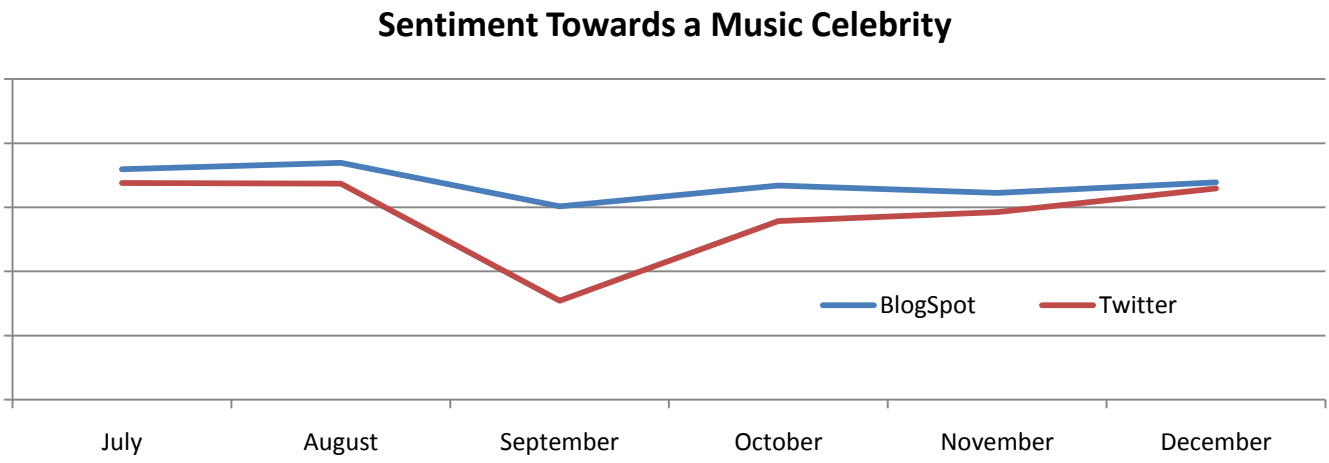
This is why weighting matters. We cannot let the voices of a group of people count more than those of other people simply because we have more data from them. If a group of people reflects 10% of the population, then their voice should count 10% of the time. We cannot allow their opinions to be more important than others.

A Practical Example

Theory is nice, but practice is best. Let's consider an example using real social media research results. For this example, illustrated in the line chart below, data has being sourced from two separate websites, namely BlogSpot (blue line on the top) and Twitter (red line on the bottom). And, we've obtained average sentiment scores towards a music celebrity over a six month period.

First, it is interesting to see that the sentiments expressed on the two websites are not the same. For some reason, sentiment, or positive feelings, towards this celebrity is much lower among Twitter users compared to BlogSpot users.

During the July and August timeframe, both Blogspot and Twitter sentiment had similarly high sentiment. However, both showed a dip in September, a dip which was much more extreme on the Twitter timeline. Sentiment became more similar again in October, and by December, it had nearly equalized.



This begs the question then, which is the right line? Is sentiment from BlogSpot users the correct answer? Or do Tweeters provide the right answer? Clearly, we cannot ignore either set of results. To get a sense of overall internet sentiment towards the celebrity, we need to combine the results together.

Combining Results

Though it might seem simple, there are many ways to create one single score that reflects the combined opinion of BlogSpot and Twitter users.

Method 1:

Calculate the average in the traditional fashion: Add up all of the scores and divide by the total number of scores.

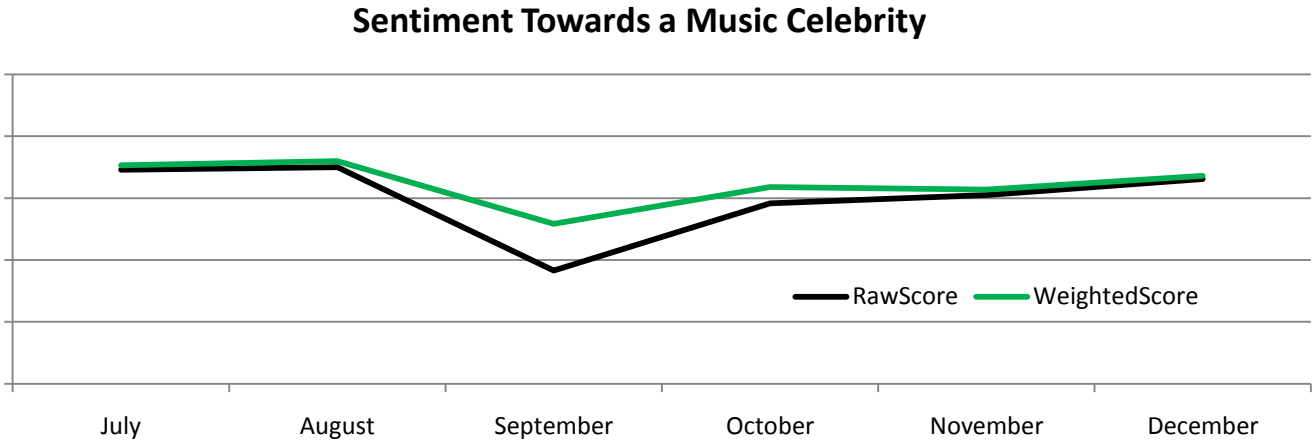
> **CON**: A few people with many opinions can overwhelm and skew the results.
>
> **CON**: Results can change from month to month if the volume from one source fluctuates every month.

Method 2:

Use a weighted average, perhaps based on the percent of internet users who visit Twitter and Blogger. Since about 11% of people visit Blogger every month, and about 5% of people visit Twitter every month, let's weight the data so that Blogger results count for twice as much as Twitter results.

> **PRO**: A few people cannot skew the results if they created most of the data.
>
> **PRO**: Trendlines reflect changes in opinions not changes in volume.

Method 1, illustrating the raw unweighted results, is identified in the line chart below by the black line (the bottom line). Method 2, which weights results based on the popularity of the website, is identified by the green line. Both methods produce similar lines that map out a dip in September. However, the dip is more pronounced for the raw unweighted trendline which over-represents Twitter opinions. The weighted results, which gives more weight to Blogger because Blogger has a larger internet presence produces a smoother line.



The Conversition Method

Conversition has developed various standardized sampling and weighting matrices, each of which reflects a specific compilation of sources to meet a variety of specific needs. Most research goals can be met using one of the standardized processes which source internet data from a variety of types of websites such as media, social networking, and blogging. Regardless of the sampling strategy employed, these processes will ensure that results are as valid and reliable as possible.

About Conversition Strategies: Conversition Strategies is a boutique online market research firm based in the US (Conversition Strategies Limited) and Canada (Conversition Strategies Incorporated). Conversition listens to consumers by applying scientific principles to the collection and analysis of social media data. Its strength lies in combining the expertise of respected market researchers with social media mavens.

www.conversition.com