An Evaluation of Methods Used to Assess the Effectiveness of Advertising on the Internet

Conducted for the Interactive Advertising Bureau

May 2010

Conducted by:

Paul J. Lavrakas, Ph.D.

Table of Contents

An Evaluation of Methods Used to Assess the Effectiveness of Advertising on the Internet

I. Executive Summary

The effectiveness of internet advertising now has been measured for more than a decade. However, there are many who remain uncertain about, and others who are dissatisfied with, whether most of the studies that measure the effectiveness of this form of advertising generate accurate (reliable and valid) results – ones that can be interpreted and applied with confidence by those for whom these studies are conducted.

Given the growing importance of internet advertising, in 2008, the Interactive Advertising Bureau (IAB) commissioned an independent study of the predominant methodologies that are used to measure the effectiveness of adverting via the internet. For the most part, these methodologies heretofore have sought to measure the branding impact of an advertising campaign by conducting a survey of consumers exposed to the campaign and consumers not exposed to the campaign. These studies include site intercept studies that sample persons in real-time as they are using the internet. It also is possible, and sometimes the case, that these studies may sample members of existing online panels.

The purpose of this report is to present the results of an objective (i.e., impartial third-party) evaluation of the reliability and validity of these research methodologies. The report contains the findings, conclusions, and recommendations of Paul J. Lavrakas, Ph.D., a former tenured professor of communication and journalism (specializing in research methodologies) at Northwestern University (1978-1996) and Ohio State University (1996-2000), former founding faculty director of survey research centers at both universities (1982-2000), and former vice president and chief research methodologist at Nielsen Media Research (2000-2007). Dr. Lavrakas was commissioned by the IAB in August 2008 to conduct this evaluation.

To conduct the study, he sought input from many executives and other practitioners within the internet advertising industry. They provided a wealth of information both verbally and in the form of print and online materials. He also used internet searches and other traditional academic methods to locate publications that were relevant to the assignment. To structure the evaluation and this report, he used two social science frameworks that address the issues of reliability and validity of social and behavioral science research studies, including those studies that strive to measure the effectiveness of internet advertising campaigns.

The IAB asked Dr. Lavrakas to conduct an objective (impartial) evaluation of the methods used by the primary research companies that measure internet advertising effectives for the online advertising industry. The IAB asked that the evaluation be straight-forward in terms of presenting an expert methodological assessment of the reliability and the validity of the predominant IAE research methods now used. The IAB also asked that recommendations be made about how to improve current approaches to measuring IAE, if in fact, improvements were judged to be needed.

The conclusions and recommendations are as follows:

1. There are many solid aspects to the manner and methodological rigor by which IAE is measured by the predominate companies currently doing these types of research studies. These include:

   - Generally robust coverage of the target population by the sampling frames that are used, in particular as it applies to the "test" groups that are sampled

   - Use of random systematic sampling, which yields a representative initially designated sample of the target population

   - Use of well-constructed questionnaires with good Construct Validity

2. However, there also are several troubling aspects to much of this research that puts the validity of the findings of most of the studies in jeopardy. These threats – which are by no means limited to IAE studies – include ones associated with the External Validity and the Internal Validity of IAE studies:

   - Their External Validity is threatened primarily by the extremely low response rates achieved in most IAE studies.

   - Their Internal Validity is threatened by the near exclusive use of quasi-experimental research designs rather than classic experimental designs.

   - Their overall validity also is threatened by a lack of valid empirical evidence that the statistical weighting adjustments used in most IAE studies do in fact adequately correct for the biasing effects of the various methodological limitations of the studies.

3. This is not to say that all the studies currently being conducted to assess IAE are reaching incorrect conclusions. Rather, it is to say that one cannot be confident whether the findings of most IAE studies are right or wrong.

4. In thinking about why the current balance of strengths and weaknesses in the measurement of IAE exists, it is important to understand that the senior researchers at the predominant companies that conduct these studies are aware of the preferred research methods that can be used to generate research findings with strong External Validity and strong Internal Validity.

5. However, the online advertising industry marketplace heretofore has neither demanded nor been willing to fund the type of IAE studies that can generate findings known to have strong Internal Validity and External Validity. Based on the research companies, publishers, and advertisers spoken to for this evaluation, *it appears that as currently perceived by most who fund IAE studies, the cost and complexity of funding studies known to be valid versus the benefits of doing this does not support the use of the more rigorous methods.*

6. To help resolve the uncertainties that currently exist, it is recommended that a series of new methodological and statistical research studies be funded *by the online advertising industry* to address the three key knowledge gaps that now exist:

   - A series of new studies should be conducted which provide a direct comparison of the findings from a classic experimental design with the findings from an otherwise comparable quasi-experimental design. This series of new methodological studies would address the unknowns associated with whether the findings from quasi-experimental designs used to measure IAE can be used with confidence to determine if exposure to an ad campaign *actually causes any of the outcomes desired by advertisers and publishers*.

- Another series of new studies should be conducted which investigate the size and nature of the nonresponse bias and other problems which may results from the extremely low response rates currently experienced by most IAE studies. This series of new methodological studies would address the unknowns associated with whether the findings from IAE studies with very low response rates have any External Validity (generalizability) beyond the proportionately small number of persons who end up providing data for the IAE studies.
- Within each of the above series of studies (those addressing quasi-experimental designs and those addressing nonresponse bias), additional analyses should be conducted to determine whether statistical weighting adjustments can reduce (or possibly eliminate) any of the biases to a negligible (i.e. ignorable) level.

II. Background and Purpose of the Study

The effectiveness of internet advertising now has been measured for more than a decade. However, there are many who remain uncertain about, and others who are dissatisfied with, whether the most common types of studies that measure the effectiveness of this form of advertising generate accurate (reliable and valid) results – ones that can be interpreted and applied with confidence, in particular by those for whom the studies are conducted.

To begin to help resolve these uncertainties and reduce the existing levels of uncertainty and dissatisfaction with how *internet advertising effectiveness* (subsequently referred to as “IAE”) is measured, the Interactive Advertising Bureau (IAB) commissioned a study in September 2008 to evaluate the reliability and validity of the predominant methods that currently are used to measure IAE. To conduct this evaluation study, the IAB hired Paul J. Lavrakas, Ph.D., a research psychologist and social science methodologist, who currently works as an independent consultant after a 30+ year career in communication and media research in both the academic sector and the private sector.[1]

The IAB directed Dr. Lavrakas specifically to focus on IAE methodologies that seek to measure the branding impact of an advertising campaign by conducting a survey of consumers exposed to the campaign and consumers not exposed to the campaign. Other means of assessing ad effectiveness (e.g., short-form questionnaires or single questions embedded within a single creative execution; correlating ad exposure with sales data; and/or click rate analysis, among others) fall outside the scope of this study. Dr. Lavrakas’s findings should not be used to validate or invalidate any methodology other than the ones he explicitly examined: i.e., long-

[1] Dr. Lavrakas is a former tenured Full Professor of communication and journalism (specializing in research methodologies) at Northwestern University (1978-1996) and Ohio State University (1996-2000), former founding faculty director of the survey research centers established at both universities (1982-2000), and former Vice President and chief research methodologist at Nielsen Media Research (2000-2007).

form questionnaire-based surveys taken by samples of exposed and non-exposed consumers recruited via either online site-intercept invitations or via an online panel.

Of note, the IAB hired Dr. Lavrakas for this project knowing full well that he had no previous experience with IAE measurement. It was explained to him by the IAB that this was done purposely, so that the evaluation would be conducted by an independent third-party with an open-mind and with no predispositions toward the topic.

Dr. Lavrakas' specific tasks were to:

1. Learn about the predominant research methodologies being used to gather and analyze data that measure IAE;
2. Assess the reliability and validity of these methods; and
3. Make recommendations about how to enhance the validity and reliability of the current state of measurement in this field.

Of special note, the larger context within which Dr. Lavrakas conducted this evaluation is one in which the marketplace generally has not demanded nor has it been willing to fund studies to measure IAE that have strong reliability and validity – studies that generate research findings that can be interpreted, and whose results can be applied, with complete confidence. That is, although the major research companies that offer IAE measurement services know how to design research studies with solid reliability and validity, too often their clients are unwilling to fund the deployment of the superior research methods and instead settle for funding studies that use research methods that have uncertain reliability and validity.

III. Methodology Used to Conduct the Evaluation Study

A. Project Chronology and Information Gathering Approaches

The approach used to conduct this evaluation was to first gather relevant information via confidential in-person meetings and telephone interviews with industry practitioners and other experts in the field of IAE measurement about the predominant methods that are used to measure the effectiveness of advertising on the internet. These meetings and interviews also led to a number of print and online materials being shared with Dr. Lavrakas in confidence. Other relevant information was gathered via internet searches for nonproprietary literature that had been published or presented on the broad topic of how the effectiveness of advertising on the internet is, and should be, measured.

The first stage of the project began in September 2008 and continued into April 2009. The second stage of the project began in July 2009 and continued into December 2009. The final stage began in January 2010 and continued into May 2010.

In the first stage of the project, Dr. Lavrakas approached the major research companies that the IAB asked him to work with in order to learn about the research methods they used to measure IAE for their clients. On the basis of the information he learned from and about these companies, he wrote a preliminary report on his findings and recommendations which was delivered to the IAB in March 2009. In April 2009 he met with the IAB and its Research Advisory Board to discuss his findings and recommendations. The IAB then distributed copies of the preliminary report to the major IAE research companies whose methods were evaluated in the report and subsequently the IAB received feedback from those companies.

After considering the feedback it received, the IAB requested of Dr. Lavrakas that a second stage of the project be conducted. The second stage began in July 2009. In this second stage, Dr. Lavrakas was asked to meet again with some of the individuals and companies he originally had met with in the first stage in order to learn about new and additional information concerning the research methods that were being used to assess IAE. In addition, Dr. Lavrakas also was asked to meet with executives of another major IAE measurement company in order to learn about the methods this company used to assess IAE. On the basis of this new information gathering effort, Dr. Lavrakas revised his preliminary report and delivered it to the IAB in November 2009. After reviewing the revised report with the IAB and receiving their feedback, he made some additional revisions and delivered a final version of the report to the IAB in December 2009. Following this, he met with the IAB and its Research Advisory Board to discuss the latest version of the report. After the meeting, the IAB sent the report to the companies whose methodologies were the subject of investigation and invited their comments.

The final stage of the project began in January 2010 with another meeting with the IAB, its Research Advisory Board, and George Ivie (Executive Director of the Media Rating Council). The purpose of this meeting was to learn and discuss the feedback Mr. Ivie had about the December 2009 version of the report. Subsequent to gaining this feedback, the IAB received further input from two of the IAE companies. Dr. Lavrakas then spoke with and exchanged emails on several occasions with senior executives of the IAB in February through May, and had several conversations and email exchanges with IAE companies to clarify his understanding of some additional information the companies had asked the IAB to have him consider in his final revision of the report. On the basis of the additional input received in 2010 from Mr. Ivie and from the IAE companies, and after a final review by the IAB of a preliminary version of the revised final report, Dr. Lavrakas produce this final version of his report in May 2010.

B. Scientific Frameworks Used to Structure the Assessment

To provide a scientific structure to the information that was gathered for this evaluation, Dr. Lavrakas used two methodological and statistical frameworks that address the validity and reliability of research in the social and behavioral sciences, including the type of research studies used to measure IAE.

- The first framework is identified with the seminal work of research psychologist and research methodologist, Donald T. Campbell, and identifies four major forms of validity: Internal, External, Construct, and Statistical Conclusion. [2]

- The second framework comes from the field of survey research and differentiates sources of bias and variance into Errors of Representation and Errors of Measurement. [3] [4] [5]

[2] Campbell, D. T. and Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research.* Boston: Houghton-Mifflin.

[3] Groves, R. M. (1989). *Survey Errors and Survey* Costs. New York: Wiley.

Together, these two frameworks provide a comprehensive perspective on all major threats to the reliability and validity of qualitative and quantitative social and behavioral science research studies.

The type of information that was gathered specifically for this IAB project addressed the following major issues and questions:

1. External Validity and Coverage Error Issues.

    a. What is the *target population* that should be studied when trying to assess the effectiveness of internet advertisements?

    b. How well does the *sampling frame* cover the target population?

2. External Validity and Sampling Error Issues.

    a. How are prospective respondents sampled from the sampling frame in order to be invited to participate in a research study to measure the effectiveness of internet advertisements?

    b. Do IAE studies use probability samples or nonprobability samples? And, what effect does this sampling design choice have on IAE study findings.

    c. How large are the *final samples* of respondents from whom data are gathered in the studies and what statistical precision do these samples sizes yield?

3. External Validity and Nonresponse Error Issues.

    a. What proportion of those who are sampled, and invited for participation in an IAE study, actually provides the data being gathered by the study? That is, what are the *response rates* for these IAE studies?

    b. What is known or can be surmised about the size and nature of the bias in these studies due to *nonresponse*?

4. Statistical Conclusion Validity, Weighting, and Adjustment Error Issues.

    a. What factors are used to adjust (weight) the data before analyses are conducted in IAE studies and how are these factors used in *weighting* the data?

    b. To what extent are the weighting adjustments used in IAE studies likely to achieve the purposes for which they are deployed (i.e., reduce or eliminate bias)?

5. Construct Validity, Specification Error, and Other Measurement Error Issues.

[4] Fuchs, M. (2008). Total Survey Error. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: Sage; pp. 896-902.

[5] Compared to Campbell's framework, Errors of Representation are related to External Validity and Errors of Measurement are related to Construct Validity and somewhat to Statistical Conclusion Validity.

   a. Which *dependent* and *independent variables* are measured in the questionnaires used to assess the effectiveness of internet advertisements?

   b. Are the questions used to *operationalize* (measure) the key variables worded, ordered, and formatted in a fashion that is reliable and valid?

   c. Are there other variables that should be included as dependent and/or independent variables?

   d. Is there any reason to expect that there is *respondent-related measurement error* that lessens the Construct Validity of the data that are gathered in these studies?

6. Internal Validity, Allocation to Treatment and Control/Comparison Groups, and Causal Inference Issues

   a. What type of *experimental*, *quasi-experimental*, and/or *nonexperimental* research designs are used to try to support the inferences that are made about whether exposure to the advertisements on the internet has caused any of the outcomes desired by advertisers?

   b. What threats exist to the validity of the causal inferences (i.e., the cause-and-effect conclusions) that are being drawn about the effectiveness of internet advertising in light of the types of research designs being used to measure such effectiveness?

   c. What *plausible alternative hypotheses* exist that may explain any observed differences in the dependent measures found between the *"test" groups* and the *"comparison" groups* in quasi-experiments?[6]

IV. Findings

What follows in this section is a detailed question-by-question presentation in which each of the 16 major questions addressed by this evaluation is described and answers and other commentary are presented.

For many of the topics that were investigated the evaluation concluded that there were not any problems of a *nonignorable*[7] magnitude associated with the reliability and validity of the findings of IAE studies.

[6] Throughout this report, when a study is contrasting a test group with an *equivalent* control group – (i.e., using an experimental design in which people have been *randomly assigned* to one or the other group, regardless of whether they also were *randomly sampled* into the study in the first place) – the term "control group" will be used to refer to the group that was not exposed to the ad campaign. In other instances where a study is contrasting a *nonequivalent* group to the test group – (i.e., using a quasi-experimental or nonexperimental design in which people were not randomly assigned to one or the other group) – the term "comparison group" will be used to refer to the group that was not exposed to the ad campaign.

[7] The term "*nonignorable*," as used in this report refers to a problem with the reliability and/or the validity of a research study that leads to inaccuracies that are of a size that makes a material difference in the conclusions that one would draw from them. The term is also used synonymously with the term, "*nonnegligible*." Use of these terms is differentiated from the use of the term, "significant," in reference to the results of a statistical test. *Significant* means that a statistical finding is reliable beyond some level of mere chance (e.g., less than the .05 level, which means

In other cases, there are ample reasons to suspect the many IAE studies are likely in error due to problems with reliability and/or validity, but ultimately that cannot be known with confidence due to a current lack of valid empirical knowledge concerning the extent to which these problems may invalidate the findings of many IAE studies.

A. External Validity and Coverage Error Issues

1. *What is the target population that should be studied when trying to assess the effectiveness of internet advertisements?*

The *target population* is the term that is used to identify the group of people to whom the findings of a research study are intended to generalize.

In the case of measuring IAE using studies that sample visitors to a website at the time they are visiting (i.e., so-called "site intercept" studies), the target population for any given ad campaign is essentially self-selected by the voluntary behaviors in which users of the internet engage.[8] Thus, it is the voluntary behavior of those people who happen to choose to visit a given website at a time that an ad campaign happens to be running, which places them into the target population of that ad campaign. And it is this ever-changing cohort that defines the target population for a given ad campaign in these types of IAE studies. For other IAE studies that may sample members of existing online panels, defining the true target population is another matter altogether, as discussed below.

Companies that advertise on the internet want to impact those visitors of the websites on which their ad campaigns will be displayed. Thus, unlike most survey situations, the fact that the target population often is "self-selected" through their own choices to visit a given internet site does not in itself add bias to an IAE study's findings.

Here, it is very important for the reader to note that this discussion concerns *the self-selected nature of the target population* and not the possible self-selected nature of the final sample of respondents that actually provides the data that are used to reflect the knowledge, attitudes, and behaviors of the target population. (The latter phenomenon is discussed below under the section on External Validity and Nonresponse Error; pp. 17-20.)

If an IAE research study randomly draws its sample from those persons who visit a website across the time period(s) the ad campaign is running, then by definition the target population for this ad campaign should be accurately represented by the *designated sample* that initially is chosen to receive (or be exposed to) an invitation to participate in the IAE study.

being certain of a statistical finding with at least 95% confidence). However, a research finding may be statistically significant largely due to having a very large sample size, even if the finding is very small in absolute size. The term "significant" is an objective term. "Nonignorable" and "nonnegligible" are evaluative (partially subjective) terms that refer to research findings that are more than merely statistically significant, in that they are judged by a research expert or other consumer of the research findings to be of a size that would change the meaning of the findings in ways that are neither negligible nor should they be ignored.

[8] It is important to differentiate the target population that actually can be sampled by an IAE study from the *population of inference* that the advertisers may want to reach with their ad campaign. The research companies that measure IAE cannot be held responsible for whether an ad campaign is being placed on internet sites where the actual population of inference that the advertisers want to reach is visiting. All that IAE measurement companies can (and should) be expected to do is sample properly from those sites where the ad campaign actually is running.

For the most part, the methodologies used by the major providers of research that sample visitors to websites to measure IAE do in fact define their target populations in this manner. Therefore, there are no adverse effects on the External Validity of the findings due to a misspecification of the target population when the research is conducted in this manner.

However, were an IAE study that is measuring a campaign being run of various websites to limit its sampling to members of an existing research panel, this sample is not likely to represent the larger target population that is exposed to the ad campaign. This is because the members of the online panel are self-selected participants in the panel and it is highly unlikely that they constitute a random (and thus representative) sample of the population that visits the websites on which the ad campaign is running.

In contrast, if the target population of an ad campaign were limited to only to members of an internet panel, then a random sampling of that membership would in fact be a valid representation of the target population. The latter might occur if an ad campaign were being pilot tested within the membership of an existing research panel. Were this approach taken, and if results on the effects of the ad campaign met the client's needs, the advertiser may then decide to place the ad campaign on internet sites that would provide far greater exposure to a larger and broader target population.

Another aspect of the question of what should be the target population for IAE studies concerns the difference between the test groups (i.e., those exposed to the ad campaign) and the control and comparison groups (i.e., those not exposed to the ad campaign) that often are used in IAE research. In the case of who is sampled and assigned to the test groups, most of the current IAE studies are a representative sample of the target population. This notwithstanding, the people who are sampled for the comparison groups in many quasi-experimental IAE studies often are not part of the target population of interest for a given ad campaign. This is because they are sampled from different websites and/or at different times than those from which the test group was sampled. In these cases, the test groups and the comparison groups are not equivalent in how well they match the target population for the ad campaign. And as discussed in more detail later in this report, despite efforts to try to make these *nonequivalent comparison groups* "equivalent" to the test groups, the adjustments that typically are made often may not accomplish what they seek to accomplish; (see pp. 20-24). In contrast, the participants in experimental ad server segmentation studies that are randomly assigned to the control group will be a representative of the target population, similar to the participants in these types of experimental studies that are randomly assigned to the test group.

For a control group to be a valid representation of the target population, the people who are sampled for an IAE study to serve in the group must be sampled at the same time and from the same websites as the test group members, with the only difference being that the control group is not exposed to the ad campaign. This currently does not happen in most IAE studies, and therefore there is a misspecification of the target population in such studies in terms of who serves as the nonequivalent comparison group.

2. How well does the sampling frame cover the target population?

Once a target population is identified, researchers must consider what *sampling frame* will be used to represent the population. Since only a very small proportion of the target population will be sampled for the research study, it is paramount that researchers are able to draw samples that accurately represent (i.e., cover) the full population of interest. If this is not done well enough, then coverage bias that is neither negligible nor ignorable may result. If this were to happen, the IAE study will not have adequate External Validity. These problems may be so severe that the entire study is rendered invalid (and thus is essentially worthless).

The sampling frame is typically thought of as a "list" of the members of the target population from which the designated (i.e., initial) sample will be selected. Depending on the mode that will be used to contact people to invite their cooperation in the research study, the sampling frame might be a list that actually exists in some physical sense or a list that could theoretically exists in some physical sense.[9]

In the case of measuring the effectiveness of internet advertisements by sampling visitors to websites, the proper sampling frame is a "list" of all those persons who visit the internet sites on which the ad campaign is being displayed during the day(s) and at the time(s) the campaign is running. As long it is this frame of people that is sampled to participate in an IAE research study, there is no *undercoverage* of the target population.

However, if a person makes multiple visits to the internet sites on which the ad campaign is running during the times the campaign is running, then a problem of *overcoverage* will exist unless corrective actions are taken. One form of corrective action is for the researchers to place a cookie onto a person's computer once s/he has been sampled so that the person will not be sampled again. But this approach does not always work as intended, because anyone who deletes her/his cookies between the time s/he is first sampled (and initially receives the cookie) and the time s/he visits the site again during the period the campaign is still running will not be detected as having been sampled twice (or more). Estimates of cookie deletion rates vary, but recent estimates by Jupiter and Belden put the rate as high as 30% of internet users, although it is uncertain how frequently cookie deleters actually do this. This 30% rate appears on the high side of what is routinely occurring and no empirical evidence was found to corroborate this level, but the actual rate is likely to be high enough to bias coverage to at least a small extent, and unfortunately for those who want to use the cookie method to help avoid overcoverage, the cookie deletion rate is likely to continue to grow. Recognizing that cookie deletion changes the probability of selection from the sampling frame, if a valid estimate of the percentage of cookie deleters in a study were known, then this percentage could be used to adjust the results of the IAE study to more accurately reflect (i.e. model) the impact of the ad campaign.

Another corrective action would be to accurately measure the number of times a person was eligible to be sampled and then use that information to later adjust (weight) the final data for these unequal probabilities of selection into the research sample. This is a routine approach taken in scientific sampling, whereby the probability of selection for every sampled person is known. (Sometimes this is known in advance of data collection and other times data are gathered during data collection to measure this selection probability.) Then it is taken into

[9] For example, in the case of those companies that build and maintain opt-in internet research panels, their sampling frame is their current list of panel members. However, since panel membership is ever changing for opt-in panels, no current list will ever exactly match the people who actually are the current members. In contrast, to conduct a random-digit dialing telephone survey of the nation, researchers do not need an actual list of all possible telephone numbers in the nation – although such a list could be created – but they do need to have all the area code and local exchange information in the nation so as to be able to randomly create any residential number that might exist in the nation.

account after data have been gathered and before the data are analyzed. Accurate measurement of the number of times a person could be sampled is not always an easy task, but even if it is not done perfectly useful data would likely be gathered simply by asking respondents to self-report the number of times they visited the websites on which the ad campaign was run during the time(s) it was run. Although for many persons the answers they provide may not be exactly accurate, the *relative differences across people* would likely provide the information that is needed to make the proper weighting corrections. In the simplest case of weighting, people might be divided into two cohorts: one cohort would be those who reported that they visited the website(s) on which the ad campaign was running at the time(s) it was running only once, and the other cohort would be those people who reported they visited the sites more than once at the time the campaign was running.

Based on the information that was able to be gathered for this report, it remained unclear how often and how thoroughly these techniques are used in current IAE studies. If none of these techniques is used in an IAE study, then coverage bias in the form of overcoverage could exist if those people with multiple chances of being sampled (e.g., heavy users of the internet sites on which the campaign is running) are differentially impacted by the ads being studied compared to those people who only visit the internet site once while the campaign is running.

In sum, there is likely to be some level of overcoverage of the target population of the sampling frame in studies that measure IAE. However, at present there is no evidence to indicate whether this overcoverage leads to coverage bias in IAE studies. Until valid information exists on this matter it seems reasonable to assume that the amount of coverage error that does exist does little to lessen the External Validity of these studies.

B. External Validity and Sampling Error Issues

3. *How are prospective respondents sampled from the sampling frame in order to be invited to participate in a research study to measure the effectiveness of internet advertisements?*

A variety of approaches can be used to draw a sample from a sampling frame. Some of these approaches use a *randomized selection process*, whereas others do not.

The sample that is initially drawn from the frame is termed the *designated sample*. Ideally, data should be gathered from all eligible members of the designated sample, although this essentially never happens in surveys, regardless of whether the surveys are for the public or private sectors and regardless of who conducts the surveys.

The least complicated, but not necessarily the most appropriate sampling scheme is a *simple random sample* in which every member of the sampling frame has an equal probability of being chosen for the designated sample. However, simple random samples can be less cost-efficient to implement and also may contribute less precision than other types of random samples.

A *systematic random sample* is one in which a fixed interval is used to sample the *n*th person from the sampling frame. This is done over and over again until the desired size of the designated sample is achieved.[10] Unlike a simple random sample that sometimes can lead to

[10] For example, the sampling interval (*n*) may be 7. The systematic sample begins with a *random start* by choosing a random person among the first *n* people. Thus, if the random start in this example is 4, then the 4th person on the

unrepresentative designated samples (e.g., it is possible, albeit unlikely, that 60%, 70%, 80% or more of a simple random sample can come from the first half of the sampling frame), a systematic sample forces sampling across the entire sampling frame in a balanced and comprehensive manner.

If information that is correlated with the major dependent variables to be studied is known about the members of the sampling frame in advance of drawing the designated sample, then that information can (and should) be used to stratify the sampling. A *stratified random sample*, when used appropriately, leads to greater precision than a simple random sample of equal size. For example, if it were known whether each member of the sampling frame was female or male, and if gender were correlated with whatever was being measured, then the sample should be stratified by gender to increase the precision of sampling (i.e., reduce sampling error for a given final sample size) and thereby possibly reduce sampling costs. However, for many IAE studies, essentially nothing is known in advance about the individual members of the sampling frame to stratify sample selection upon. Thus, stratified random sampling is not an option for most of the IAE studies that currently are conducted by sampling visitors to websites.

In contrast, IAE studies that may be based on sampling from an existing online research panel easily lend themselves to stratified sampling, and thus this should be used routinely. Companies that build and maintain research panels have myriad data about their panel members. Thus, if an IAE study were to use members of an existing research panel, the sampling should be stratified on those characteristics that are known (or are assumed) to correlate with the dependent variables of interest (e.g., likelihood of purchasing the product or service being advertised).

In sum, for most IAE studies, the major research companies that study IAE use a systematic random sampling design to initially select the visitors to a website on which an ad campaign is running to participate in a study of that campaign. This constitutes the sampling design for forming their designated samples. Given that they do not have variables on which they can stratify this sampling, the systematic approach that is used is the best one to control sampling error.

In contrast, whenever it is appropriate to test an ad campaign among members of an existing research panel, a stratified sample should be used, as such samples are the most efficient and effective ones to reduce sampling error.

4. *Do research studies that try to measure the effectiveness of internet advertising use probability samples or nonprobability samples and how does this effect Sampling Error and External Validity?*

For a research study to be able to quantify the level of precision associated with its findings – what is often termed *sampling error* – a *probability sample* must be used. For a sample to be a probability sample, the researchers must be certain that each member of the sampling frame has a non-zero chance of being selected and they must know the probability of selection that is associated with each member of the sampling frame. For example, if there will be 1,000,000 visitors to a website during the times an ad campaign will be running and if the researchers

sampling frame is the first member that is sampled; the 11th person (4 + n = 11) is the second member sampled; the 18th person (4 + 2n = 18) is the third sampled, and so on.

evenly spread the sampling across all the times and select 10,000 of the visitors to be invited to participate in the ad effectiveness research study, then each of the members of the sampling frame has been accorded a chance of selection equal to 10,000/1,000,000 or 0.01 or 1 in 100. The fact that every member of the sampling frame has a *known nonzero chance of selection* is what makes this a probability sample. Of note, the probability of selection need not be equal for each member of the sampling frame in order for a sample to be a probability sample. Instead, what is required is that each member of the sampling frame must have a *known nonzero* chance of being selected.

If the probability of selection for each member of the sampling frame is unknown, and/or if some members of the sampling frame have zero chance of selection, then the research study is said to have a *nonprobability sample*. When using a nonprobability sample, the size of the sampling error should not be computed as the computation is statistically meaningless. That is, an uninformed researcher may use a sampling error formula with data from a nonprobability sample, but the results of the calculation have no statistical reliability or validity.

Most IAE studies that sample visitors to websites draw samples that at least approximate probability samples, including those that use a test-control/comparison design to study internet advertising effectiveness. This is because most of these studies try to use a sampling procedure that gives everyone who visits a website at the time the study is being conducted an equal chance of selection. However, as discussed previously, this breaks down for certain visitors who come to the website(s) more than once while the ad campaign testing is being conducted. As also mentioned previously, this can be corrected if the researchers measure how often the visitor was eligible for selection. If they know this information about all of their sampled respondents, then they can assign a unique probability of selection to each person and thereby their sample will remain a probability sample. If they do not gather or otherwise know this information, or if they do not use a procedure that assures everyone has an equal probability of being sampled, then they no longer have a probability sample. And in those instances, their ability to know with what confidence their statistical findings generalize beyond the sample that provides data is essentially zero, because they cannot calculate a meaningful sampling error for their study.

An IAE study that were to draw its sample from members of an existing online panel can readily deploy probability samples as it is easy for the researchers to know the probability of selecting any given member of the panel. In most instances these types of IAE studies would sample panel members with an equal probability of selection. In these studies sampling error is readily calculated and the researchers can quantify the level of confidence associated with any of their statistical findings, *but only as it generalizes to the nonrandom opt-in panel from which the sample was drawn.* That is, in these types of IAE studies, sampling error cannot properly be applied to any larger target population to which an ad campaign may be oriented, since the opt-in panel itself was not selected via a probability sample from that larger target population.

5. *How large are the final samples of respondents from whom data are gathered in IAE studies and what precision do these probability samples sizes yield?*

Once a random probability sampling approach has been chosen to select members of the sampling frame for study, a decision must be made regarding the number of respondents from whom data must be gathered to meet the desired statistical needs of the IAE study. It is this *final sample size* that determines the size of the sampling error associated with a given IAE research study and its particular probability sampling design.

And, it is the size of the sampling error that determines the size of the confidence intervals for the IAE study's findings and thereby directly affects the precision of the statistical tests that will be performed with the data that are gathered to help determine whether or not the ad campaign was effective.

If all the members of the sampling frame who were selected for participation in an IAE research study provided the data the researchers were seeking from them, then the size of the final sample would equal the size of the designated sample. However, this is essentially never the case in any survey (not just IAE studies) due to *survey nonresponse* (which is discussed below in detail in the following section; pp. 17-20). Therefore the final sample size often is many increments smaller than the size of the designated sample, and therefore the designated sample size must be many increments larger than the final sample size.

Since sampling error is predicated on the final sample size, the researchers must determine the final sample size they need for their statistical purposes once all the data have been gathered, and then use that number to extrapolate to the size of the designated sample with which they must start in order to achieve the necessary final sample size. For example, if the final sample sizes needed are 500 for both the test and comparison groups in an IAE study, and if only 1 in 50 sampled persons provide data (i.e. a two percent response rate), then the designated sample must be 50 times the size of the final sample, or 25,000 for both the test and comparison groups.

As discussed below, nonresponse is extremely high in almost all IAE studies currently being conducted that sample visitors to websites. Because of this, the designated sample sizes that are required for most of these types of IAE research studies are at least 20 times larger (and often 100 or more times larger) than the final sample sizes needed. In contrast, if an IAE study were to sample from an existing research panel, nonresponse within the panel would be comparatively low and thus the designated sample size may only need to be two or three times the final sample size.

Currently, final sample sizes used in most IAE studies that use probability samples fall into the 500-2000 range. This range provides precision in the plus/minus 2 to 4 percentage point range with a 95 percent level of confidence. This range of sample sizes provides more than adequate statistical precision for the major purposes to which these IAE studies are put – i.e., making broad decisions about whether or not an internet ad campaign was effective. In instances were the sample size is at the lower end of this range (e.g. < 800) and the clients want subsample analyses to be conducted (e.g., break out upper-income females and males from lower income females and males), these subsamples may not have enough members in them to provide precise analyses. Thus, subsample analyses based on small sized subsamples (< 100 in the subsample) will have relative large sampling errors and thereby may lack External Validity (and Statistical Conclusion Validity).

In sum, in the vast majority of cases, the sampling approaches used by the predominant companies that measure IAE with a probability sample are appropriate and have the precision needed to draw reliable statistical conclusions about the data that have been gathered.

C. External Validity and Nonresponse Error Issues

As noted above, survey nonresponse exists in essentially all surveys and nowadays generally is regarded as one of the two greatest challenges to survey accuracy.[11] Thus, the problems discussed in this section are by no means limited to the realm of IAE measurement.

Nonresponse occurs when complete data are not gathered from all the eligible persons that have been sampled for participation in a research study. Nonresponse can exist at the *unit-level*, whereby no data whatsoever are gathered from a sampled person or at the *item-level* whereby a sampled person does provide some of the requested data but fails to provide data for one or more of the variables being measured.

6. What proportion of those who are sampled and invited for participation in an IAE study actually provides the data being gathered for the study?

As it relates to IAE studies, nonresponse may occur at more than one stage of the research depending on whether the sample for the study is taken from “live” visitors of internet sites on which the ad campaign is running or whether the sample comes from members of an existing research panel.

For those IAE studies that sample visitors to websites at the time an ad campaign is running, nonresponse takes place at the time invitations to participate in the study are presented to the visitors.

For those IAE studies that may sample members of existing online panels, nonresponse takes place at the time they are invited to participate in the study *and* at a prior time when the panel was being built. Thus for example, 80 percent of those online panel members invited to participate in a particular IAE study may do so, but that does not mean that the response rate of this study is 80%. Instead, the true response rate is a miniscule (and typically unknowable) number which is a combination (i.e., the product, in a multiplicative sense) of the response rate in forming the original panel and the response rate for the particular IAE study within the panel. Thus even though the response rate among panel members for a particular study may be very high, the true response rate is likely to be extremely low given that response rates for building all opt-in internet panels are extremely low.

In most surveys, the vast majority of unit-level nonresponse is due either to *noncontacts* or to *refusals*.

In the case of surveys that are conducted to measure IAE, a noncontact would result if a person did not receive, or simply failed to notice, an invitation that was sent to her/him to participate in the survey for which s/he was sampled.

Most studies now being conducted to measure IAE send invitations to their sampled respondents at the very time the person is visiting the website on which the ad campaign is running. If the person is using a browser or other software that interferes with the invitation

[11] Presently, the other greatest challenge is the use of unrepresentative sampling frames, with considerable coverage biases, such as (a) those used for opt-in internet panels (aka “access panels”), that do not adequately represent (cover) the target population of interest, and (b) those telephone surveys that only use traditional list-assisted RDD landline frames that nowadays provide coverage of less than 70% of the U.S. population..

reaching her/him then this will result in a noncontact. That is, since the sampled person never knows s/he has been invited to participate, s/he becomes a nonrespondent due to noncontact. If the invitation does come to the sampled person but s/he he fails to notice it (e.g., it simply does not catch her/his eye as s/he is quickly scanning part of a webpage), this is another form of nonresponse due to noncontact. The size of noncontact-related nonresponse in the surveys conducted to measure IAE is unknown, but it is reasonable to estimate it to be less than 10% of those who are sampled.[12]

Experience shows that *refusals are the primary source of nonresponse* in most surveys, including those that are conducted to measure IAE. This occurs when a sampled person "notices" (even if simply for a millisecond) the invitation to participate in the survey but chooses to ignore it or otherwise not to cooperate.

There are many reasons that people refuse to cooperate in surveys, but the reasons most likely to explain the bulk of the refusal-related nonresponse to IAE surveys are:

- Lack of interest;
- Lack of time;
- No incentive, or an inadequate incentive, being offered by the researchers; and
- Concerns about confidentiality and privacy.

Taking into account all the reasons for nonresponse to IAE surveys, the amount of nonresponse to these surveys is massive. The response rates that almost all IAE studies garner are less than 5% and in many cases are far less than 1%. This means that 19 persons out of every 20 (and oftentimes more than 99 out of every 100) that are sampled to participate do <u>not</u> provide any data.

This is a serious threat to the External Validity of these studies if the very large group of persons who are sampled but from whom no data are gathered differ in nonignorable ways on the variables of interest from the proportionally very small group of sampled persons who do participate in these types of surveys. And, as discussed below, there are ample reasons to suspect that such differences do exist between responders and nonresponders.

7. *What is known or can be surmised about the size and nature of the bias in most IAE studies due to nonresponse?*

As noted above, nonresponse is extremely large in almost all studies that strive to measure IAE. And, the vast majority of nonresponse is due to a *self-selection* process that plays out at the level of the individual who is sampled from the sampling frame but refuses to cooperate. (Of note: The reader is reminded that nowadays almost all surveys suffer from large amounts of nonresponse, not just those surveys that are used to measure IAE; although many other surveys achieve response rates far higher than is common in most IAE studies.)

[12] The estimate is based on (a) on the proportion of persons that are thought to use advanced software to block the predominate mode used to send survey invitation in internet ad effectives studies and (b) the fact that these invitations are readily noticed by the vast majority of those to whom they are sent.

For many reasons, the likelihood that sampled nonresponders, as a group, would provide data that are essentially identical to the data that are provided by the IAE study's responders, as a group, appears to be low. And, due to the amount of the nonresponse in most IAE studies, even if differences between the responders and nonresponders were relatively small, the size of the *nonresponse error* is likely to be large enough to be nonnegligible and thus nonignorable.

Furthermore, *even if sophisticated weighting techniques are used to adjust the data to try to account for nonresponse, currently there is no valid empirical evidence available to show that these adjustments when used in IAE studies can and do in fact eliminate (or even reduce) the bias that nonresponse may be causing.* (Of note, and as discussed in detail later, a major recommendation of this evaluation is that the online advertising industry band together to fund studies to close this important knowledge gap by gathering valid empirical evidence about this issue.)

In addition to these concerns, what more can be surmised about the nature of the likely nonresponse bias in most IAE studies?

To try to answer this question, it is instructive to revisit a very well conceived and very well executed study that the IAB commissioned in 1997, along with Millward Brown, to investigate online advertising effectiveness.[13] This study garnered a response rate of 45% to the first stage of the research from among the approximately 80,000 website visitors that were sampled to participate, and a response rate of 47% to the second stage of the research; overall this yielded a response rate of 21% for those sampled respondents who completed both stages of the research (final sample *n* = 16,758).[14]

Nonresponse bias was not a topic often addressed in the mid-1990s, but to their credit the authors of this 1997 report explicitly addressed the issue of possible nonresponse bias in their methodological appendix.

Their reasoning holds relevance for the current evaluation of how IAE now is being measured more than a decade later:

> First, does the propensity to cooperate in a survey that tries to assess IAE correlate directly (positively) with the propensity to be affected by exposure to internet advertisements? If it does, then it is those persons most likely to participate in such studies that also are most likely to be impacted by the ads the surveys are meant to study. Logic suggests that this is very likely to be happening. However, even if this is true, it does not mean that current IAE studies are wholly incorrect in concluding that ads are bringing about at least some of the positive effects desired by advertisers. *But it does suggest that most of the current studies are likely to be overestimating the size and strength of these effects because of the nonresponse bias associated with their final samples.*
>
> Second, and under a worst case scenario, what conclusions would be drawn about nonresponse bias if all of the nonresponders to the current IAE surveys were totally

[13] Briggs, R. et al. 1997. *IAB Online Advertising Effectiveness Study.* New York: Internet Advertising Bureau and Millward Brown Interactive.

[14] Achieving response rates such as those in this IAB study would be much more costly today than it was in 1997.

> "impervious to [internet] advertising"?[15] If this were the case, then even that does not completely cancel out all of the positive effects that current IAE studies are showing. Rather, it likely dilutes most of them. But logic dictates that the actual amount of this dilution is less than the extreme case where there is no positive impact whatsoever.

In sum, due to the likely nonresponse biases that exist, most IAE studies may be overestimating the positive effects of the internet ad campaigns that are being measured. However, there presently exists no valid empirical basis on which to estimate by exactly how much or how often this is happening. (Of further note, and as discussed in detail later, a major recommendation of this evaluation is that the online advertising industry also band together to close this important knowledge gap by funding studies to gather valid empirical evidence about this issue.)

D. Statistical Conclusion Validity, Weighting, and Adjustment Error Issues

Traditionally, there are two primary reasons that survey data should be weighted (adjusted) before analyses are conducted:

- The first is to correct for any unequal probabilities of selection of the designated sample from the sampling frame.
- The second is to try to correct for the possible biasing effects of survey nonresponse.

As noted previously, in order for proper weighting to be done to correct for unequal probabilities of selection, researchers must know what the selection probability is for each of the members of the sampling frame. If this individual-level probability is not known then corrections for this factor cannot be made properly. So for example, any *nonprobability sample,* in which there is not a known probability of selection for every member of the designated sample, cannot be adjusted validly. Depending on the size of this problem, the final statistical conclusions of such a study may be rendered wholly invalid.

For proper weighting to be done to try to correct for nonresponse, the values of the variables in the survey dataset known about the respondents (e.g., percentage female, percentage 18-24 years of age, percentage Hispanic, etc.) must be compared to values of these same variables in the target population; (these population values often are termed *population parameters* or *universe estimates*). For nonresponse adjustments to reduce any nonresponse bias that may exist in the unweighted survey data, the variables that are used in this stage of weighting must correlate with the key dependent variables the research is meant to study. However, although using variables that correlate with a study's key dependent variables for weighting is a necessary condition for reducing nonresponse bias, *it is not a sufficient condition* that automatically results in elimination, or even a reduction, of nonresponse bias. For the researchers to have confidence that the weighting they perform to adjust for the effects of nonresponse does in fact work, they need to have an independent and valid external source of evidence to point to, such as a valid nonresponse bias study (see pp. 39-43).

Unfortunately, neither of these types of weighting adjustments can be made with confidence in the current studies used to measure IAE. This includes the studies that sample people as they

[15] Briggs, R. et al. 1997. *IAB Online Advertising Effectiveness Study.* New York: Internet Advertising Bureau and Millward Brown Interactive; p. 77.

are visiting websites on which an ad campaign is running and studies that may draw samples of members of an existing internet opt-in panel to use in a test of an ad campaign or others types of empirical investigations of the impacts of an ad campaign.

Weighting for Probability of Selection. In the case of making adjustments for unequal probabilities of selection in studies that sample people who are visiting the websites on which an ad campaign is running, IAE researchers do not know with certainty that their cooperating respondents all were sampled with an equal probability. One of the reasons for this is that in the current methods being used, researchers most often do not know with reasonable certainty how many times someone has visited the websites during the times the campaign was running. This happens because of the cookie deletion problem and how it changes selection probabilities. This leads to unequal probabilities of selection because those who have visited the sites more than once during the campaign may not get blocked from being invited to partake in the study more than one time and the researchers currently do not routinely gather data from respondents to measure how many times they might have been invited to participate in the study. All this leads to the problem of unadjusted overcoverage.

There is not a uniform agreement among industry experts about whether this overcoverage is leading to coverage bias in IAE study data. One prominent researcher does not believe that it is.[16] Regardless, a positive step could be taken to address this potential source of bias by routinely gathering information from each person who participates in an IAE study about the number of times s/he could have been sampled during the time the ad campaign was running. This information then could be used to weight the final samples to help adjust for these unequal probabilities of selection. Furthermore, this approach entirely circumvents the cookie deletion problem.[17] Granted the data that respondents provide about their frequency of visiting websites may not be completely accurate in an *absolute sense*, but it seems very reasonable to assume that it is quite accurate in a *relative sense,* and it is the relative difference between study participants that needs to be adjust for, not the absolute difference. Gathering such data will add slightly to study costs, but clients need to recognize that the value of being able to properly weight for unequal probabilities of selection justifies the small incremental cost.

For IAE studies that may sample members of existing panels, researchers often would not need to perform weighting adjustments for unequal probabilities of selection because all the panel members sampled for a given study likely would be sampled with equal probabilities of selection. In rare cases where the sampling probabilities of different panel members were not equal, the researcher would know the selection probabilities for each sampled panel member.

Weighting for Nonresponse. In the case of making adjustments for nonresponse, the researchers who are conducting studies of people visiting websites while an ad campaign is running *do not have all of the population parameters about the larger target population they would need to adjust their data to reflect.* Thus, this remains a problem (a) as long as the size of the nonresponse in these studies remains as large as it now exists or (b) until a valid series of industry-funded nonresponse bias studies were conducted that happened to provide evidence that nonresponse is not a major source of bias in these types of IAE studies. As discussed later

[16] Havlina, B. 2008. "Impact of Cookie Deletion on Research Results." New York: Dynamic Logic.

[17] A potential problem still remains when cookie deletion is not detected or avoided, because someone who is sampled more than once might participate in the study more than once. However, the chance of this happening is so low that there is no reason to believe this is having any appreciable biasing effects in IAE studies.

in this report (pp. 39-43), the latter certainly is a topic for which new and valid industry-sponsored research needs to be conducted.

Adjusting for nonresponse in IAE studies that may sample members from existing internet panels is a simple process since the companies that build and maintain these panels know a great deal about their members and thus sampled responders to a given study can easily be compared to the sampled nonresponders. *However, this form of weighing only will adjust for nonresponse within the context of the panel itself, and not for the massive amount of nonresponse that occurred at the time the panel was being formed.* Since it is this latter type of nonresponse that threatens the representativeness of almost all opt-in internet panels, making nonresponse weighting adjustments within the context of the panel may not make any sense. However, if the internet panel membership already is weighted to try to reflect the characteristics of some known larger target population the panel purports to represent (e.g., all English language adults in the United States), then it may make sense to perform the additional nonresponse weighting adjustments within the sample of responders to adjust how they differ from the larger panel from which they were sampled.

8. What factors are used to adjust (weight) the data before analyses are conducted in IAE studies and how are these factors used in the weighting?

The variables that typically are used for weighting in IAE studies are gender, age, income, and internet usage. In addition, select psychographic variables that are linked to the focus of each specific study also may be used.

However, the purpose of this weighting and the manner by which it typically is done in IAE studies deviates wholly from the traditional approach that often is used to make nonresponse weighing adjustments.

As noted above, the predominant research companies that measure IAE via sampling visitors to websites on which the ad campaigns are running most often do not have data on the relevant population parameters of their target populations. Without such data, they cannot make valid comparisons between these relevant characteristics of the proportionately small subset of responders who provide data in their studies to the same characteristics among members of the sampling frame. Without such population parameters, nonresponse cannot be addressed properly by weighting.

For example, assume an advertising campaign is being tested that is promoting a service that has special appeal to introverted personalities. If a research company knew that 40% of the visitors to websites running a particular ad campaign were introverts (i.e., 40% of the target population was an introvert), but that only 20% of those who responded to the survey measuring the effectiveness of the ad campaign were introverts, then this information could be used to weight the final data set by this psychographic characteristic. However, since the companies do not know anything with confidence about such characteristics of the target population for a particular ad campaign, valid nonresponse adjustments cannot be made.

Apart from the two traditional approaches to weighting is used by the predominant companies that measure IAE, there is another way that weighting is used in the IAE studies that employ a test/comparison group design.

This other approach to making statistical adjustments to IAE data before they are analyzed, concerns the "*equivalency*" of the test and nonequivalent comparison groups that typically are used in these studies. Of note, if a true experimental design were used in these studies – one in which all visitors to websites running the ad campaign, regardless of when they visited the site, are randomly assigned to either an equivalent control group or a test group – theoretically there would be no need to do this weighting. However, because a classic experimental design is rarely used in IAE studies, the visitors assigned to the comparison group and to the test group are not equivalent, and thus weighting is used to try to compensate for this.

The manner in which this weighting it done is to use demographic variables, such as gender, age, and income, that are gathered from both the test and comparison groups; and also may include using some psychographic variables gathered from both groups that are linked to the specific nature of the ads being measured. Through a series of iterations (know as "raking" or "rim weighting"), the composition of the comparison group on these demographic and psychographic variables is adjusted to make the comparison group resemble the test group *on these characteristics*.

However, it is very important to understand that these adjustments do not address the issues of External Validity (discussed previously in this report) or of Internal Validity (discussed later in this report, pp. 28-35).

In sum, there are no weighting adjustments that currently are made in IAE test/comparison group studies (ones that do not use random assignment to the two groups) that provide any certain improvement in the External Validity or Internal Validity of IAE research studies.

For those rare experimental ad server segmentation studies that randomly assign existing panel members to the test or control group – whereby the test group is shown the online ads and the otherwise equivalent control group is not – weighting for nonresponse is not necessary in order to gain strong Internal Validity, but weighting could be beneficial to enhance the External Validity (generalizability) of the study if it were known what respondent characteristics correlated with their propensity to participate in the study and if these characteristics also were correlated with the dependent variables in the study.

9. To what extent are the statistical weighting adjustments used in IAE studies likely to achieve the purposes for which they are deployed (i.e., reduce or eliminate bias)?

This issue addresses the questions of how well, if at all, weighting adjustments improve the accuracy of findings compared to (1) the findings from an unweighted dataset that is not fully representative of the target population and (2) the findings from a dataset that would not require weighting adjustments because it already was fully representative of the target population (i.e., a "gold standard" dataset)?

For weighting to improve the accuracy of any IAE study findings, the variables that are used to weight must be correlated with the dependent variables of interest and with the mechanisms that caused the unweighted sample to not be fully representative of the target population. In the case of IAE studies there are three major mechanisms that may lead to unrepresentative samples: (a) coverage error, (b) nonresponse error, and (c) lack of random assignment to test and control conditions. Although IAE researchers may know information about how well the weighting variables correlate with the dependent variables in their studies, for many reasons they do not know much about how the weighting variables correlate with the aforementioned mechanisms that led to the unrepresentative samples in the first place.

The issue of how well weighting can correct for unrepresentative samples is a controversial one throughout the survey and marketing research industries. There are those who appear to make the blanket claim that such procedures do in fact work well enough to have confidence in the final weighted results essentially in all circumstances.

In contrast, the traditional view within the behavioral and social sciences is a much more cautious and skeptical one. Weighting does do what is expected of it under many circumstances, but certainly not all. The circumstances in which it is not likely to attain what its proponents claim for it, such as those discussed above, are those with which IAE researchers too often are faced: i.e., circumstances in which the researchers simply cannot be confident that the correct variables have been used to base the weighting upon.

Ultimately, the burden of proof that weighting is working as it is intended lies with the researchers who make the claims that it is: (as noted in the movie *Jerry McGuire*, "Show me the money!"). And, in the absence of reliable and valid empirical evidence that a specific approach to weighting is in fact improving the accuracy of the final results, there is no reasonable basis to have confidence that it does do this.

E. Construct Validity, Specification Error, and Other Measurement Error Issues

Construct Validity essentially refers to the issue of whether or not the researchers actually measured what they claim to have measured with the variables they used in a research study. In the case of measuring whether or not an internet ad campaign achieved it goals, the researchers ultimately are trying to measure whether the campaign changed the knowledge, attitudes, behavioral dispositions, and/or behaviors among those exposed to the campaign. If the actual data that are gathered in an IAE study are not valid measures of these key constructs then the study will suffer from a lack of Construct Validity.

In studying the effectiveness of an internet ad campaign, there are certain key dependent and independent variables that must be gathered to allow the researchers to provide findings that their clients want to know regarding the effects of the ad campaign. The *dependent variables* represent measures of which effects, if any, the campaign had on those exposed to the ad(s). The *independent variables* that are gathered allow the researchers to (a) weight the dataset, if weighting is used, and (b) analyze the dependent variables in more depth; for example, to investigate differences in the dependent variables related to the respondents' demographic and psychographic characteristics.

If a study does not gather the correct dependent variables – which is referred to as *specification error* – or if the study does not measure variables in a reliable and valid manner – which is referred to as *questionnaire-related measurement error* – the resulting data would have little, if any, Construct Validity and thus would not serve the purpose for which the study was conducted. Were this to be the case, the study would not provide valid information about whether or not the ad campaign was effective.

10. Which dependent and independent variables are measured in the questionnaires used to assess the effectiveness of internet advertisements?

The predominant research companies that study the effectiveness of internet ad campaigns typically gather the following types of variables in their questionnaires, adapted to the specific ad campaign being studied:

- Dependent Variables
    - Brand awareness
        - Unaided recall
        - Aided recall (may display an ad or ads from the campaign to jog memory)
    - Brand exposure
    - Brand Image
        - Valance (positive, neutral, negative)
        - Compared to brand's major competitors
    - Future propensity to purchase brand
    - Actual brand purchases
- Independent Variables
    - Demographics
        - Gender
        - Age
        - Income
        - Place of residence
    - Lifestyle and other Psychographics
        - Use of the Internet per week
        - Use of a shared or non-shared computer
        - Past purchasing behavior
        - Other attitudes about the domain of products or services into which the brand falls

All of these variables are important to measure in IAE studies and it is entirely appropriate that the current IAE studies routinely include such variables.

11. Are the questions used to operationalize (measure) key variables worded, ordered, and formatted in a fashion that is reliable and valid?

This issue addresses whether or not there is likely to be nonnegligible *questionnaire-related measurement error* introduced into the data due to a poorly constructed questionnaire.

The examples of questionnaires that have been used in actual IAE studies that were made available for review for this evaluation allowed for an assessment of whether there is reason to expect error due to the questionnaires that are used in these studies.

After evaluating the wording of the questions, the ordering of the questions, and the formatting (layout) of the questions, there were no problems found to suggest that the questionnaires create any appreciable measurement error, either in the form of variance or bias. That is, the wording of the questions, the ordering of the questions, and the layout of the questions used by the predominant IAE measurement companies essentially follow best practices in survey questionnaire construction.

However, minor suggestions could be made about how to reword some of the questions that are asked to improve their reliability and the value of the data they gather, but none of these suggestions are of major consequence to the reliability and validity of the data that are gathered via these IAE questionnaires.

In sum, the questionnaires that are used to gather data in IAE studies are not contributing Measurement Error and thereby are not lessening the Construct Validity of these studies.

12. Are there other variables that should be included as dependent and/or independent variables?

A review of the research literature during the past two decades on the topic of how IAE can be measured uncovered a few additional dependent variable constructs that do not appear to routinely be measured in current IAE studies, but can be recommended for inclusion. [18] [19] [20] [21] [22]

These include:

[18] Haugtvedt, C. P and Priester, J. R. (1997). Conceptual and methodological issues in advertising effectiveness: An attitude strength perspective. In W. W. Wells (Ed.) *Measuring Advertising Effectiveness*; pp. 79-93. Hillsdale, NJ: Erlbaum.

[19] Briggs, R. and Hollis, N. (1997). Advertising on the web: Is there response before click-through?, *Journal of Advertising Research*, 37(2), 33-45.

[20] Pavlou, P. and Stewart, D. (2000). Measuring the effects and effectiveness of interactive advertising: A research agenda. *Journal of Interactive Advertising*, 1(1). http://www.jiad.org/article6.

[21] Rodgers and Thorsten, E. (2000). The interactive advertising model: How users perceive and process online ads. *Journal of Interactive Advertising*, 1(1). http://jiad.org/article5.

[22] Li, H. and Leckenby, J. D. (2004). Examining the effectiveness of internet advertising formats. In D. W. Schumann and E. Thorson (eds.), *Internet Advertising: Theory and Research,* pp. 203-224. Hillsdale, NJ: Erlbaum.

- Certainty of intention to purchase a brand
- Trust in the brand, both in terms of credibility and benevolence
- Site visit experience, in terms of the reactions the respondent had to the brand website
- Does the ad elicit any "feedback" from the respondent, which can be considered a prerequisite of "interactivity"

The following are independent variables that do not appear to be used routinely in IAE studies, but can be recommended for inclusion:

- The number of times in the past 30 days a person visited a website on which an ad campaign is running
- The number of times during the ad campaign a person visited a website on which the ad campaign was running
- The number of times a person deleted cookies on her/his computer during the time period during which people were being invited to participate in an IAE survey
- Race and/or ethnicity
- Educational attainment
- Home owner/renter status
- Presence of non-adult children in the home

In sum, there are some additional dependent and independent variables that researchers should (re)consider using in future IAE studies. Adding some or even all of these variables should not adversely affect survey response rates.

13. Is there reason to expect nonnegligible respondent-related measurement error that lessens the Construct Validity of the data that are gathered in these studies?

Respondents can be a troubling source of survey measurement error if they are unable or unwilling to provide accurate and complete data.

If respondents are asked for information they do not know, they may, for various reasons, "make up" answers rather than report they do not know an answer. If respondents are asked for too much information or information that causes them too much effort to report – both of which would be an example of a high *respondent burden* – they may *satisfice* by coming up with the "easiest" (but not necessarily an accurate) answers they believe will appear credible to the researchers.

In reviewing the questionnaires used by the predominant companies that conduct IAE studies, there was no indication that the questions that are asked are difficult or otherwise taxing for respondents to answer.

Nor is there any indication that too many questions are asked thereby causing undue *respondent fatigue*. In fact, the questionnaires appear to be ones that take less than five minutes to answer and are filled with items that have little or no *cognitive complexity*.

In sum, there is no indication that respondent-related measurement error is limiting the Construct Validity of the studies that currently are conducted to measure the effectiveness of internet ad campaigns.

F. Internal Validity, Allocation to Treatment and Controls Groups, and Causal Inference Issues

Possibly the most important lesson that this IAB evaluation project has identified is the extent to which the validity issues that are addressed in this section of the report are not well understood, and thus their importance is not fully appreciated, by the on-line advertising industry.

If the on-line advertising industry is serious about the importance of having sound (reliable and valid) research on which to base decisions about whether a given advertising campaign brings about any of the changes in knowledge, attitudes, and behaviors that the advertisers are seeking, then a major shift in attention to, and understanding of, the issues of “Cause-and-Effect Validity” – which the social science research literature refers to as Internal Validity – will be required.

This major shift also should be accompanied by willingness within the advertising industry

- to acknowledge its failure in the past to require and to provide the funding necessary so that valid research designs are used to assess the impact of advertising campaigns, and

- to band together to fund a series of valid methodological research studies to compare the validity of experimental research designs with that of quasi-experimental and nonexperimental designs in the measurement of the effectiveness of internet advertising.

Only by engaging in a muscular investigation of these validity issues will the on-line advertising industry know with confidence whether the quasi-experimental and nonexperimental methods, which predominate the types of IAE studies that currently are conducted, can provide “valid enough” findings to give advertisers what they need in order to determine what worth a given ad campaign may have.

Ultimately, advertisers should want to know with confidence whether their ads are causing the effects they want to achieve, such as increases in brand awareness, loyalty, and purchasing.

For this, they need research with a high degree of *Internal Validity*, which means that the research design that is used to measure the impact of the ad campaign provides strong support for *cause-and-effect conclusions* (e.g., X causes Y) to be drawn with confidence. An example of a cause-and-effect conclusion is, “Exposure to the ad campaign led to an increase in the intention to purchase Brand UVW of 24%.” (In this example, the ad campaign is X, the independent variable, and the intention to purchase the brand being advertised is Y, the dependent variable.)

Anyone has the right to speculate on the causal implications of research findings regardless of what type of research design is used – and many researchers and nonresearchers do this routinely.

However, it is only through the use of a rigorous *unconfounded* experimental design – also referred to as a “true experiment” or a “classic experimental design” – that researchers can interpret the causal implications of their findings with great confidence.

Of note, it was just this type of research design that was used in the 1997 IAB Online Ad Effectiveness study, and it is instructive to re-examine the research objectives of that experimental study:[23]

> *“Achieving the study objectives required the methodology to meet a number of key requirements:*
>
> 1. *Measure, with high degree of accuracy, the impact of Web [ads] within the actual environment where Web [ads] operate.*
>
>   *This requires [an experimental] test that involves:*
>
>   - *Real Web users (sampled during the natural course of surfing) who view…*
>   - *Real [Web ads] (Which are part of actual online campaigns) that are located on…*
>   - *Real media properties (with an established audience and market) during the…*
>   - *Normal course of Web media consumption.*
>
> 2. *Isolate the effects of the [ad] exposure from all other factors which impact a consumer’s relationship with the brand.*
> 3. *Use measurements that can be administered via a survey which will accurately gauge the relationships that individuals have with brands and advertising.”*

It is Objective 2 (above) that indicates the need for an true experimental design, as it is only through a classic experiment that the effects of exposure to an ad campaign can be *isolated “from all other factors”* that may impact a consumer’s knowledge of, attitudes towards, and behaviors with the product or service being advertised and studied.

In contrast, when an experimental design is not used, researchers (and others interpreting the findings of the research study) are entitled to exercise their *professional judgment in speculating* about what the causal inferences of the study’s findings may be, but they have no scientific basis to claim that the research study design provided internally valid results – ones that can be interpreted with confidence that “X caused Y.”

[23] Briggs, R. et al. 1997. *IAB Online Advertising Effectiveness Study.* New York: Internet Advertising Bureau and Millward Brown Interactive; p. 71.

Put another way, when a classic experimental design is used, it is the methodological power (rigor) of the design itself that “speaks” with confidence as to whether or not there are statistically significant (i.e. reliable) causal effects.

In contrast, when an experimental design is not used, it is the researcher’s own judgment that speaks to whether there may be meaningful casual effects. This is not to imply that a researcher’s professional judgment about whether causal effects exist always is in error in the absence of using an experimental design. *Rather, it is to say that without an experimental design, the researcher cannot point to internally valid empirical findings to support her/his contention that the campaign is or is not working.*

Unfortunately, this scientific reality does not stop many from drawing definitive causal inferences, when instead they should be exercising more caution. That this occurs as often as it does appears in good part due to too many of the people who are funding quasi- and nonexperimental research studies – e.g., advertisers and other clients – having business roles and responsibilities that do not require an adequate understanding or appreciation of the strengths and limitations of various research methods. And, because they do not have the scientific understanding or appreciation of why such cause-and-effect claims may not have strong internal validity, they too readily accepted and interpret findings from quasi- and nonexperimental research studies with more confidence than should be accorded to them.

The characteristic of a true experiment that provides its power and the rigor for drawing causal inferences with great confidence is <u>random assignment</u>.

A true experimental research design requires a test group and a control group that are equivalent in all ways except for the researchers’ manipulation of the experimental treatment. In the case of measuring IAE, the “treatment” is the ad campaign that the advertiser hopes will bring about desired effects. *For a research study of advertising effectiveness to be an experiment, the persons who are studied must be individually assigned in a random fashion to either a control group or a test group.* This random assignment occurs after people have been sampled (be that via random sampling or non-random sampling), but before anyone is exposed to the treatment.

Here it is paramount that readers understand that *random assignment* is <u>not</u> the same as *random sampling*.

Random assignment in an experiment takes place after sampling has been conducted – be that random sampling or nonrandom sampling.

Random sampling is related to whether an IAE research study is likely to generate externally valid findings that can be generalized beyond the study participants with confidence.

In contrast, random assignment is related to whether an IAE research study is likely to generate internally valid findings that can be interpreted with confidence that “X caused Y,” where X is exposure to an ad campaign and Y may be an increased tendency to purchase the brand being promoted by the campaign.

Furthermore, the External Validity and the Internal Validity of a research study exist independently of the other. That is, a study may have: (a) strong External Validity without any Internal Validity; (b) strong Internal Validity without any External Validity; (c) neither strong Internal nor strong External Validity; or (d) both strong Internal and Strong External Validity.

What sponsors of IAE research studies should be willing to fund are studies that have both strong External Validity and strong Internal Validity.

14. What type of experimental, quasi-experimental, and/or nonexperimental research designs are used to try to support the inferences that are made about whether exposure to advertisements on the internet has caused any desired outcomes?

There basically are two types of research designs that are used by the predominate companies that measure IAE by sampling visitors to websites on which an ad campaign is running.

One is a classic experimental design with equivalent test and control groups. These studies have come to be known in some circles as “ad server segmentation studies,” and appear to represent a small minority of the studies that currently are conducted to measure the effectiveness of internet advertising campaigns. There are senior researchers on the staff of the predominant companies that measure IAE who know how to design these “gold standard” experimental studies.

That so few of these studies are conducted appears in large part due to the possible ignorance and/or short-sightedness of advertisers regarding the importance of funding IAE studies that use an experimental design and/or the unwillingness of advertisers to provide the incremental funding necessary to conduct experimental studies.

The other type of design is not a classic experiment, although in most instances it can be regarded as being *quasi-experimental* in that it has test and comparison groups, even though the two groups are *nonequivalent*.

When these quasi-experimental designs are used, which constitute the vast majority of the studies that are currently conducted to measure IAE, researchers typically apply various statistical adjustments to try to make their test and comparison groups more “equivalent.” Unfortunately, and despite the researchers’ efforts and good intentions, no one can know with confidence how well these statistical weighting adjustments actually counteract (i.e., offset) the effects of the initial nonequivalency of the test and comparison groups.

Why this is of crucial importance to the validity of the cause-and-effect interpretations that are made with these quasi-experimental IAE studies, is that if one cannot be confident that a test group and a comparison group are *equivalent in all important ways* – with the exception that the test group was exposed to the ad campaign and the comparison group was not – then there may be other explanations (so-called *plausible rival alternative hypotheses*) apart from the presence or absence of the ad campaign exposure that account for some or all of any observed differences between the test and comparison groups.

Schematically, this means that if X is the ad campaign and Y is a change in the test group’s level of brand loyalty, then when using a quasi-experiment even with statistical weighting adjustments, one cannot be certain that there is not some other unidentified factor(s), Z, that correlate(s) with X and that actually is(are) causing the observed change in Y (brand loyalty level).

If so, then the causal relationship between X and Y is *spurious* (false) and therefore really not causal at all. If this were the case in an IAE study that does not use an experimental design, then even though the results may suggest (and be interpreted with confidence by some) that the ad campaign has “caused” the effects the clients were seeking, it may be an incorrect conclusion. For IAE studies that may sample existing online panel members, the same validity concerns exist depending on whether an experimental design is used or not.

Concerning IAE studies that may sample online panel members, experimental designs should be used routinely, because essentially they should not cost any more to conduct than quasi-experimental and nonexperimental IAE studies that sample panel members. That is, an existing online panel is ideal for testing the effects of an online advertising campaign using a true experimental design. Setting up the experiment is no more complicated than *randomly assigning* panel members to either the test or the control group. The test group is exposed to the ad campaign and the control group is not. Then both groups are asked the same set of questions and their answers are compared to determine whether exposure to the campaign brought about any of the effects desired by the advertisers. Preferably the test group’s exposure occurs via a means with *mundane realism* – i.e., a means that simulates the experience of what one encounters while surfing a website. But even if it does not, the experiment can yield valuable feedback about the power of the ad campaign. That is, in the simplest of these types of experiments, the randomly assigned test group is asked merely to look at an online advertisement. The randomly assigned control group is not shown the advertisement. The two groups are then asked the same questions about the brand. If there are no differences in the responses from the two groups, the advertisers can conclude the ad is a failure and go about revising or scrapping that particular ad campaign. In other words, if an ad campaign fails to lead to any positive effects under this most sterile of research testing it is unlikely to yield desired effects in “the real world.”

15. What threats exist to the validity of causal inferences that are being drawn about the effectiveness of internet advertising in light of the types of research designs being used to measure such effectiveness?

Experimental Designs. Regardless of whether an IAE study samples from visitors to websites or from members of an existing online panel, whenever a classic experimental design is used to study the effects of an internet advertising campaign, the test and control groups start out being equivalent because of the random assignment of each sampled respondent to one or the other of these groups. By randomly assigning a person to either a test or control group, the only differences between the two groups should be whatever the researchers do to expose the test group to the ad campaign under study. In contrast the control group is not exposed to the ad campaign, but rather is exposed to “neutral” alternative content or to no additional content at all.

However, even when using an experimental design the study does not always get conceptualized or implemented as it should, nor do things always go as planned and expected. For example, an experiment may be unintentionally *confounded* by the occurrence of some factor or event that changes the test group’s or control group’s knowledge, attitudes, or behaviors towards the service or product under study, apart from the ad campaign content.

For example, if the researchers inadvertently communicate something to the test group (but not to the control group) that sensitized the test group to why the study was being conducted, that difference in itself may make the respondents in the test group more prone to report data that are favorable to the brand being studied. In this case, the information about why the study was being conducted would confound the experimental treatment, as the ad campaign no longer would be the sole difference between the test and control groups.

Another potential threat to the Internal Validity of an experiment is the possibility that over the passage of time, the test and control groups may have different experiences outside the confines of the experiment with whatever is being studied. However, since almost all of the experimental studies conducted to measure IAE gather data very soon after the respondents are exposed to either the test ad campaign or to the neutral control content, the possibility of this "history" threat is greatly diminished or essentially eliminated. Were this not the case, then the researchers would need to measure additional exposure to the brand that might have occurred after the initial exposure to the ad campaign (or to the neutral content) but before the time the data for the study were gathered. Of note, a "welcomed" complication can arise where initial exposure to an ad campaign may "change" the test group members in a way that subsequent exposure to the brand outside the confines of the experiment further enhances the test group's knowledge, attitudes, and/or behaviors to the brand. In this case, the advertisers are benefiting beyond the direct/immediate effects of the ad campaign through this indirect effect that is enhancing the effects of subsequent exposure to the brand. This is not a confounding influence on the experiment. Furthermore, the potential for indirect effects can be studied within the context of the experiment as long as data are gathered from respondents about their subsequent level of exposure to the brand after their initial exposure via the internet ad campaign.

Quasi-Experimental Designs. There are myriad ways that test and comparison groups may not be equivalent in a quasi-experimental design – one that does not randomly assign persons to a test or control group – and some or all of these ways may lead to spurious conclusions about cause-and-effect relationships despite any statistical weighting adjustments the researchers may make to try to increase the "equivalency" of the two groups.

As discussed previously in this report, statistical weighting adjustments often are made in IAE studies to try to offset the nonequivalence of the test and comparison groups on certain factors. *That these adjustments are considered necessary attests to the initial nonequivalence of the two groups (regardless whether random* sampling *is deployed).*

The factors for which the two groups are adjusted typically include gender, age, income and internet usage, and also may include psychographic measures that are linked to the domain of services or products into which the brand being studied fits. The rationale for, and the assumption underlying, these adjustments is that (a) by making the comparison group "look" more like the test group in terms of gender, age, income, internet usage and possibly for some psychographic characteristics, then (b) the two groups are "equivalent enough" to not allow their remaining "nonequivalencies" to account for any differences that may be observed between the test and comparison groups in the key dependent variables being studied.

For example, imagine that unadjusted data from one of these quasi-experimental studies showed the test group had 25% more awareness of the brand after the ad campaign was run than did the comparison group. But, then imagine that after the "equivalency" adjustments were made the test group now was found to have only 10% more awareness of the brand than the comparison group. However, it is possible that there are other important factors on which the two groups differ but for which adjustments were not able to be made. If so, then had the researchers known to, and been able to, make these additional adjustments the test group may have been found not to display any heightened awareness of the brand over the comparison group. Granted, this example is mere speculation, but it is a possibility that needs to be considered whenever a quasi-experimental design without random assignment is deployed in an IAE study.

16. What plausible alternative hypotheses exist that may explain any observed differences in the measures being taken between the test group and the comparison group in a quasi-experiment that samples visitors from websites?

As explained above, whenever a quasi-experimental research design is conducted, the nonequivalency between the test and comparison groups makes it uncertain whether any observed differences in the dependent measures between the two groups (e.g., a relative increase in intentions to purchase the brand among the test group) can and should be attributed to the impact of the treatment to which the test group was exposed (e.g., the content of the ad campaign) or if there are other plausible explanations for the observed effects.

The following are plausible alternative explanations for why the quasi-experiments used to measure IAE that show positive impacts on the test group may not mean it was the ad campaign that brought about some or even any of these observed changes.

Nonresponse as a plausible alternative explanation. A plausible alternative explanation for why some quasi-experimental IAE studies may generate spurious findings is related to the differential effects of nonresponse. As discussed previously, the propensity to respond to an IAE study, on average, is very low among those who are sampled for such studies. That is why response rates to the studies that sample visitors to websites are generally less than 1%; i.e., less than 1 in 100 persons sampled for the test and comparison groups actually participates. If those persons sampled for the test group who are most likely to respond are also most likely to be impacted by the advertising to which they are exposed, then data produced by this group will overestimate the positive effects that are measured. This does not necessarily mean that the campaign did not have any impact on those exposed to it, but the size of the true impact of the campaign would not be measured accurately because of the effects of nonresponse.

Frequency of Visiting Websites as a plausible alternative explanation. Another plausible alternative explanation for positive findings about IAE that come from these quasi-experimental designs is the difference between the test and comparison groups in the frequency with which they visit the website(s) from which they were sampled for the quasi-experimental study. If members of the test group on average are relatively heavy visitors of the websites on which the ad campaign is running, whereas the comparison group on average are less heavy visitors, this nonequivalency between the two group could lead to an overestimate of the impact of the ad campaign because, for example, the test group would have had more chances to be exposed to the campaign. The higher frequency of visiting certain websites among the test group also may be correlated with their being more interested in advertising on these websites than a comparison group that visits the websites less frequently. This too would confound any interpretation about how much, if at all, the ad campaign had any of the effects desired by the advertisers. Using the frequency with which a respondent visited the websites on which the ad campaign was running during the times it was running as a weighting variable (or as a covariate) may somewhat correct for these “nonequivalencies,” but there is no guarantee that it will do so completely.

Timing of Visiting the Websites as a plausible alternative explanation. Another plausible alternative explanation for the observed effects of an internet ad campaign could be that the test and comparison groups differ in terms of the timing when they were measured within the duration of the campaign. In these quasi-experiments, the test and comparison group members typically are not being sampled simultaneously. Because of these time differences in sampling, each of the groups may have nonequivalent experiences with the website or with other factors, all of which could affect the data they provide.

In sum, there very well may be plausible alternative explanations for some of the positive effects that are observed in quasi-experimental IAE studies other than the inference that it was the ad campaign that should be credited for bringing about all of the observed effects.

V. Conclusions

As detailed above, there are many solid aspects to the manner and methodological rigor by which IAE is measured by the predominate companies currently doing these types of research studies.

However, there also are a number of troubling aspects to much of this research that puts the validity of the findings of most IAE studies in question.

These threats include ones associated with the External Validity and Internal Validity of IAE studies. Their External Validity is threatened primarily by low response rates. Their Internal Validity is threatened by the near exclusive use of quasi-experimental designs rather than classic experimental designs with random assignment to test and control conditions.

In thinking about why the current balance of strengths and weaknesses in the measurement of IAE exists, it is very important to acknowledge that most of the senior researchers at the predominant companies that conduct these studies appear to be aware of the preferred research methods that can be used to generate research findings with strong External Validity *and* strong Internal Validity.

But as currently perceived by most advertisers and other clients, the cost of funding valid IAE studies versus the benefits of doing this appears not to support the use of these more rigorous methods.

Another troubling aspect of most current IAE studies is the heavy reliance on statistical weighting adjustments to try to correct for limitations inherent in the method of sampling, the very low response rates, and/or the nonequivalency of test and comparison groups. There is no solid empirical evidence about how well these adjustments in IAE studies accomplish what they are intended to accomplish. And in fact, the research literature to date suggests that such adjustments regularly fall short of their goals.[24] [25] Thus, until such reliable and valid evidence is available anyone interpreting the findings of an IAE study that uses weighting adjustments to correct for methodological limitations in the study design or its execution, is duly warned to be cautious in not over-interpreting the findings.

[24] Chang, L. and Krosnick, J. (2009). National Surveys via RDD Telephone versus the Internet: Comparing Sample Representativeness and Response Quality. *Public Opinion Quarterly*, 73(4), 641-678.

[25] Baker, R. et al. (2010). AAPOR Report on Online Panels. http://www.aapor.org/AM/Template.cfm?Section=AAPOR_Committee_and_Task_Force_Reports&Template=/CM/ContentDisplay.cfm&ContentID=2223

A clear implication of these conclusions is that clients of IAE studies should be more careful when interpreting and applying the results of any study that does not use a classic experimental design. These clients should strive to better understand the limitations of the quasi-experimental designs they are funding and in the weighting procedures that are routinely used in many IAE studies. By better understanding these limitations regarding the validity and reliability of most of the studies they now are funding, clients hopefully will be better informed to (a) use the results of these quasi-experimental studies with greater caution, (b) more often choose to fund studies that deploy a classic experimental design, and (c) be willing to help fund new methodological research to improve the current state of knowledge regarding the validity and reliability of quasi-experimental IAE studies and all IAE studies that have extremely low response rates.

In line with this last point, new methodological research – aka "research on research" – should be conducted to determine with much more precision than currently exists what are the limitations to the validity and reliability of (a) the current quasi-experimental designs being used, (b) online ad effectiveness research studies with very low response rates, and (c) the weighting adjustments that commonly are used in most IAE studies. The nature of this new research is addressed in the following section on Recommendations.

Until such new methodological research is conducted, the IAE industry will continue to use the current methods that have become accepted within the industry, but as noted above the findings from quasi-experimental studies and all studies with very low response rates and those with weighting adjustments need to be used cautiously until more is known about the validity of such IAE studies.

VI. Recommendations for Much-Needed Research on Research

As previously noted, it is possible that the error (bias and variance) in the vast majority of current studies that measure IAE is negligible and thus ignorable.

Unfortunately, no one has valid and reliable evidence to answer whether the amount and nature of the error is ignorable or not.

The major threats to the validity and reliability of most of the current online advertising effectiveness studies are three-fold:

- The quasi-experimental designs with nonequivalent test and comparison groups, despite the statistical adjustments that are made to "equivalentize" the groups, may be yielding findings with inadequate Internal Validity.

- The extreme low response rates experienced with the recruitment methodologies currently used – which contributes to the specter of an appreciable amount of nonresponse bias in the study findings – may negate the External Validity of the studies.

- The uncertainty associated with whether the weighting adjustments that are routinely used to correct for the methodological limitations in most IAE studies do in fact improve the accuracy of the final results, and if so, by how much.

Until these issues are answered authoritatively, the validity and reliability of all the studies that use quasi-experimental designs to study IAE and/or all the IAE studies that have low response rates and/or those that use weighting adjustments will remain unknown.

Therefore, in order to close these conspicuous and critical knowledge gaps, the following recommendations are proposed.

A. *A series of well-funded new research studies should be conducted which allows for the direct comparison of the findings from a classic experimental design with the findings from an otherwise comparable quasi-experimental design. This series of studies addresses the unknowns associated with whether the findings from quasi-experimental designs used to measure IAE can be used with confidence to determine if exposure to an ad campaign causes any of the outcomes desired by advertisers. As part of this research, and if the quasi-experimental results are not equivalent to the experimental results, then further analyses should be conducted to determine if weighting the quasi-experimental results can make them equivalent to the experimental results.*

More than a decade ago, the IAB funded a very well conceived and well executed classic experiment to investigate the effectiveness of online advertising.[26] This study provided internally valid evidence that the ads that were tested at that time had a number of positive impacts on the groups of persons who were exposed to them versus the groups of persons in the unexposed, but otherwise equivalent, control groups. The use of this type of research design provided evidence that could be very confidently interpreted as strongly supporting the cause-and-effect conclusions that were drawn from that study.

It may be that some or many or even most of the findings from the current quasi-experimental studies that nowadays are being conducted to measure IAE are valid; or are at least “valid enough” for the purposes to which the clients want to put the findings.

The problem is that even with the statistical adjustments that are made to try to reduce the nonequivalence between the test and comparison groups in these quasi-experimental studies, one cannot be certain that the interpretations drawn from the studies about the possible effects that were “caused” by the ads are in fact accurate.

As discussed previously in this report, it may be that these effects are spurious (false) and are instead explained by some other (undetermined) factors than the exposure of the test group to the online ads compared to the nonexposure of the comparison group.

To close this critical knowledge gap, a series of *paired-studies* should be conducted that compare the findings of a classic experiment design with that of a quasi-experiment design similar to those that commonly nowadays are used to measure IAE.

Of note, there are four factors that would need to be in place to make this line of industry-sponsored paired-study research appropriate (valid) for the critical need it will serve:

[26] Briggs, R. et al. 1997. *IAB Online Advertising Effectiveness Study.* New York: Internet Advertising Bureau and Millward Brown Interactive.

- There should be at least three of these paired-studies conducted, and preferably more, so that the findings from each can be used to determine (via triangulation) whether there is any consistency across the series of study findings.

- Each paired-study should be conducted either by (a) an independent organization (i.e., not one of the companies currently measuring IAE) that is known to do valid and high quality experimental research or (b) by two of the companies that currently are measuring IAE, but with one of the companies conducting the classic experiment part of the paired-study and the other conducting the quasi-experimental part of the paired-study. Having any company that currently is in the business of IAE measurement conduct both parts of one of these paired-studies would be an ill-advised conflict of interest for that company.

- Other than the use of a classic experiment to *randomly assign* sampled website visitors or existing panel members to either a test or control group and a quasi-experiment to sample nonequivalent test and comparison groups, the two studies within each pair should be identical in all other aspects of their research methods. This includes the look and content and type of invitation sent to recruit participants, the questionnaire and questions used to gather the dependent and independent variables, and the approximate sample sizes of the test and control/comparison groups used in each of the paired studies.

- If any of the paired studies is conducted by two companies currently measuring IAE (with one company doing the experimental study and the other company doing the quasi-experimental study), then each company should privately write up their results and submit them to an interactive advertising industry oversight group, which then would allow the group to determine whether the findings from the quasi-experimental part of the study are essentially the same as or different from the findings from the experimental part of the study.

An example of a paired-set study would be one in which the exact same ad campaign is used in an experimental design and in a quasi-experimental design. The exact same sampling scheme would be used for each design (e.g., a systematic random sample). The same websites and times/days would be used (although depending on the quasi-experimental design, there may be some differences here in sampling comparison group members). The same exact types and content of invitations to participate in the survey would be used in each design. The exact same questionnaires, in terms of format, question wording, and question ordering would be used in each design. Finally, the same independent variables (e.g., gender, age, income, number of times visiting the websites) should be measured and used either as covariates or as adjustment factors in each design. The only factor that would differ would be that the test and control groups in the experiment would be formed via true random assignment, whereas the test and comparison groups in the quasi-experiments would be formed via the nonequivalent procedures now used for most IAE studies.

In evaluating the results of the series of paired studies, the findings from the experimental designs would serve as the “gold standard” against which the findings from the quasi-experimental designs are compared. To the extent the findings of the quasi-experimental designs mirror those of the experimental designs, the industry can be confident that findings from high quality quasi-experimental studies of IAE have adequate Internal Validity and can be used to draw cause-and-effect conclusions with confidence. To the extent the findings for the quasi-experimental designs do not mirror those of the experimental designs, the industry will be duly forewarned to carefully rethink what future value such quasi-experimental studies may have in trying to determine IAE.

Furthermore, if the quasi-experimental results do not mirror those of the experimental study, then additional analyses should be conducted to investigate whether there may be statistical weighting adjustments that can be applied to the quasi-experimental data to bring those findings "close enough" into line with the experimental findings so that any remaining differences between the two study designs would be deemed negligible (and thus ignorable). If this were found to be possible, then the interactive advertising industry would learn a very important lesson about how to use quasi-experimental studies so that their weighted results are in fact known to be valid.

In sum, only by conducting a series of paired studies like this will advertisers, publishers, IAE research companies, the IAB, and others concerned about the validity of IAE measurement come to learn whether the quasi-experimental approaches most often used to conduct IAE studies have adequate Internal Validity to justify their continued use.

B. A series of well-funded new research studies should be conducted which investigate the size and nature of the nonresponse bias and other problems which may results from the extremely low response rates currently experienced by IAE studies. This series of studies addresses the unknowns associated with whether the findings from IAE studies with very low response rates have any External Validity beyond the persons who end up providing data for the studies. As part of these studies, analyses should be conducted to determine if there is an optimal way to weight the data that are gathered in IAE studies so as to reduce the amount of nonresponse bias that may be present to a negligible (i.e., ignorable) level.

As noted in the Findings Section of this report, the response rates for most of the studies that currently are conducted to measure IAE are less than 5%; and in many instances they far are less than 1%.

Granted, a low response rate, in and of itself, does not assure that there will be an appreciable amount of nonresponse bias in a research study. However for IAE studies in which 19 out of 20 sampled respondents (i.e., a 5% response rate) or 99 out of 100 sampled respondents (i.e., a 1% response rate) do not provide any data, there very likely is some nonignorable biasing factor or factors that explain why certain types of sampled respondents do not provide any data. That is, it is highly unlikely that those who do provide data in an IAE study with an extremely low response rate are a random (and thus representative) subsample of all the potential respondents who were sampled for the study in the first place.

Currently there exists no valid empirical evidence to inform the online advertising industry about the size and nature of the possible nonresponse bias in IAE studies. This leaves a considerable and conspicuous knowledge gap about whether the studies have any External Validity whatsoever.

If they do not, then results of these studies cannot be generalized beyond those relatively few sampled persons who provided the data. Furthermore, no one knows how much more funding per study would be required to generate data that would not be invalidated by nonresponse bias.

To close these crucial knowledge gaps, a series of nonresponse investigations and a series of nonresponse bias studies must be conducted.

There are two primary purposes for these studies:

- ***Find cost-effective means for raising response rates to IAE studies***

  A number of factors impact the decision a sampled respondent makes in whether or not s/he will participate in a research study. The factors that appear most relevant to raising the response propensity of those sampled for IAE studies are:

  - ***How likely an invitation will be noticed and read***

    If the intended recipient is not aware that s/he has been sampled to participate in a research study, there is zero-probability that s/he will do so. From what has been learned in conducting this evaluation, it does not appear that the industry knows with confidence what portion of those who are sampled to participate in an IAE study (one that samples visitors from websites) never realize they have been sampled, because they either never notice the invitation or, if they see it, they never read it. Since DHTML invitations appear not to be blocked (making it unlikely that an invitation will not be noticed), new experimental research should be conducted that focuses on the "look" of these types of invitations to determine whether there are design features that elicit appreciably higher response rates.

  - ***The substance and language used to make the invitation***

    Oftentimes "short is sweet," but other times "too short" is not. In terms of inviting people to join an IAE study, there is not enough known about how these invitations should be worded and formatted. For example, a best practice in the survey literature is to use polite wording, such as the word "please." However, few of the invitations reviewed for this evaluation project made any concessions to explicit politeness.

    Furthermore, confidentiality and concerns about privacy have been consistently found to affect the likelihood of some types of people to participate in surveys. However, none of the invitations reviewed for this evaluation project addressed the issues of confidentiality and privacy. New experimental research should be conducted that uses various wordings and various looks (e.g., a casual look, a professional look, a "hot" look, etc.) for the wording that is used to make the request and that proactively mention something about confidentiality and privacy.

    For IAE studies that might sample members of existing panels, the response rates would be expected to be substantially higher than for the studies that sample visitors to websites. But even these types of IAE studies would benefit from investigating how the substance and language used to make an invitation to participate in a given study might be improved.

  - ***On whose behalf the invitation comes and by whom it is sent***

    The research literature indicates that the sponsor of a research study and the group conducting the study can impact the response propensity of those who are sampled to participate. New experimental research should be conducted that varies whether or not this information is included in the invitation and if so, what level of prominence is it accorded.

- ***The rewards (incentives,) if any, that are offered for someone to participate***

  Incentives often are a “last resort” means of trying to motivate cooperation among those sampled respondents for whom other means of motivating them to participate in a research study have failed. To the extent that those who are motivated primarily or solely by incentives differ in their reaction to internet advertising, bringing them into the final samples of IAE studies in greater proportions will lower the bias their absence is creating. Noncontingent incentives have been consistently shown to outperform contingent incentives of even greater value. Using a small noncontingent incentive (e.g., valued at $1 or $2) and a larger contingent incentive (e.g., valued at $5) is likely to be the most cost-effective incentive approach to raising cooperation and compliance to an invitation for an IAE survey. This notwithstanding, new experimental research should be conducted to generate empirical findings about the effect of incentives on response rates in IAE studies.

  Only with such valid evidence can advertisers and IAE researchers know whether the incremental costs that incentives would add to IAE studies would be beneficial to their interests in gathering valid information about IAE.

All of these factors and possibly others should be investigated in a series of experimental studies in which *factorial designs* are deployed to learn what effect the various factors have on response rates (and at what cost). Using *unconfounded full factorial designs* for these experiments will allow for a valid causal interpretation of the main effects that are being tested, as well as for testing for possible interaction effects the factors may have on each other.

It is only through the use of well conceptualized and unconfounded experimental studies that the interactive advertising industry will learn with confidence how response rates to IAE studies can be raised and at what cost.

- ***Investigate at what levels of nonresponse nonignorable nonresponse bias exists in IAE studies and whether there is a cost-effective way to correct for it***

As noted above, low response rates, in themselves, do not necessarily mean that a research study will be invalidated by nonresponse bias.

However, the extremely low response rates that commonly occur in current IAE studies, in particular those that sample visitors to websites, – be they experimental or quasi-experimental – are suggestive that nonignorable nonresponse bias may be present in their findings. Furthermore, this bias may not be eliminated, or even appreciably reduced, by the statistical adjustments made to IAE studies.

Thus, a series of nonresponse bias studies should be conducted to allow the online advertising industry to better understand what is the size and nature of the bias due to nonresponse in current IAE studies and what, if anything, can be done to reduce its effects in lowering the validity of the study findings.

There are several ways that nonresponse bias can be investigated, and each of the following types of studies should be considered in deciding how to gather data to address this issue:

- ***Follow-up a subsample of nonresponders after the original study is completed and compare them to responders***

  In this type of study a subsample of nonresponders to the original study are recontacted at some point after the original study is completed and asked the same demographic, psychographic, and brand-related questions the responders answered when they responded to the original study. The sooner a follow-up study is conducted after the original study ends, the better. The larger the response from the original nonresponders in this follow-up study, the greater will be the confidence in the generalizability of the comparisons that are made between the characteristics of the responders versus the original study nonresponders who respond to follow-up study. If the differences are minor (negligible) this provides credence to the conclusion that nonresponse bias in the original study may well be ignorable. If the differences are major (nonnegligible) this provides evidence that the at least some of the findings of the original study are likely to be invalid.

- ***Incent a subsample of initial nonresponders with ample rewards to gain cooperation from a large proportion of them during the time of the original study***

  If a person is sampled to participate and receives an invitation to do so, but does not respond, that does not mean that further efforts to gain cooperation from this person must be abandoned. In fact, *refusal conversions* – efforts to convince initial nonresponders to cooperate – are a routine best practice in many surveys.

  For example, within some relatively short time of seeing the initial invitation for an IAE study, but not responding to it (5-20 seconds later), the person could be sent a different-looking second invitation that includes an incentive offer. This incentive would certainly motivate some of the initial nonresponders to respond; and, the more attractive the incentive, the larger the proportion of initial nonresponders who then will respond. In this way, a study of IAE could have both initial responders and some of the initial nonresponders in the final sample. Then these two groups could be compared as to whether they gave similar or dissimilar data to the questionnaire used to judge the effectiveness of the ad campaign.

  Ultimately, what is of greatest interest is whether the same conclusions about the ad campaign's effectiveness would have been drawn using only the initial responders, as would be drawn using both the initial responders and the initial nonresponders who eventually responded after being offered the incentive. This would provide valuable evidence as to whether the data from only the initial responders suffers from nonignorable nonresponse bias.

- ***Conduct a formal nonresponse bias study that achieves a high level of response from both former responders and former nonresponders***

  The most elaborate (and expensive) way to study nonresponse bias is to mount a formal study which samples both former responders and former non-responders. A very robust methodology must be used to gain a high degree of cooperation (preferably >60%) from both groups. Studies like this have shown that former responders are very likely to cooperate again when given modest incentives, but that former nonresponders require much greater incentives to gain their cooperation. Once the cooperation of sampled former responders and former nonresponders is secured, they are asked to complete the same questionnaire. In the case of studying nonresponse bias in IAE studies, part of the questionnaire could involve exposure to a new ad campaign followed by questions about it, similar in nature to the types of questions typically asked in IAE studies. In this way, the data from former responders and former nonresponders can be compared to determine whether the two groups differ in any nonignorable ways in their reactions to the advertisement(s). By inference, if there are no important differences, this would be evidence to suggest that the findings from the original study were not invalidated by nonresponse bias in that study. In contrast, if the two groups did differ in important ways, this would be evidence that at least some of the original findings are invalid due to nonresponse bias.

Of note, the nature of response/nonresponse is likely to differ quite a bit in important ways depending on whether the IAE study is one that samples visitors to websites or one that samples existing online panel members. That is, not only will the size of the nonresponse between the two types of samples differ considerably, but the factors that underlie why a sampled person does or does not respond are also likely to be considerably different, depending on which type of sample is used in the IAE nonresponse study.

All of the above approaches to studying nonresponse bias lend themselves to statistical modeling that likely would yield information about the size and nature of nonresponse bias in IAE studies under various assumptions about the size and nature of the nonresponse.

In combination with these various suggestions about how to study nonresponse and nonresponse bias in IAE studies, analyses should be conducted to investigate whether the effects of any nonignorable nonresponse bias can be reduced (or possibly eliminated) through the use of weighting adjustments to the data. But for this to be possible, the nonresponse bias studies need to be conducted first, as their findings will help identify the “gold standard” against which a weighted IAE study dataset would be compared. If such weighting adjustments were shown to work “well enough,” then the interactive advertising industry would learn a very important lesson about when to discount the results of studies with low response rate and when not to be concerned by low response rates.

In sum, by using a series of nonresponse bias studies, such as those described above, the online advertising industry could begin to be informed about the size and nature of any nonresponse bias that might be found via these methodological investigations. It then may be possible to devise reliable and valid statistical adjustments that could be applied to future IAE research data to provide future researchers with a valid empirical basis on which to adjust their data for nonresponse bias.

APPENDIX A

Persons Who Generously Provided Information Used in this Evaluation

Keith Berkelhamer, *Turner Broadcast System*

Rex Briggs, *Marketing Evolution*

Josh Chasin*, comScore*

Jim DeMarco, currently at iVillage, formerly at *Time-Warner*

Gerald Dirksz, *comScore*

Jon Gibs, *Nielsen Online*

Bill Havlina, Ph.D., *Dynamic Logic*

Molly Hilsop, *InsightExpress*

Anne Hunter, *comScore*

George Ivie, *Media Rating Council*

Stephen Jepson, *InsightExpress*

Matthew Katz, *Turner Broadcast System*

Tom Kelly, *SafeCount*

David Kudon, *Turner Broadcast System*

Drew Lipner, *InsightExpress*

Dan Murphy, *Univision*

Keith Nielsen, formerly at *DoubleClick*

Rory O'Flynn, *InsightExpress*

Mary Ann Packo, *Dynamic Logic*

Jean Robinson, *Dynamic Logic*

Christine M. Winnicki, *Time-Warner*

Stephanie Young-Helou, *InsightExpress*