Managing Deceptive Respondents in Online Surveys

Alex Kalamarides, Ph.D., Managing Director, and Eric Shuster, CEO, IntelliClear, Inc.

January 14 2020

Abstract

The modern era of online quantitative studies has brought an onslaught of unqualified survey takers motivated by incentives, sometimes rich ones, and wreaking havoc for market research professionals. Several algorithms for removing such deceptive respondents have been successfully employed. These algorithms are especially successful for removing those with distinctive or unusual response patterns, but they can fail with deceptive respondents whose responses look statistically similar as those of valid ones. Thorough removal of such intractable deceptive respondents (IDRs) through screening can be prohibitively expensive and time consuming. However, a surprisingly robust proportion of IDRs can be safely tolerated in the survey without materially affecting survey results, representing a reasonable and viable alternative to extensive or exhaustive screening measures and pointing at simple rules for practitioners in view of relevant economic trade-offs.

Introduction to Deceptive Respondents

Quantitative market research surveys consist predominantly of close-ended – oftentimes, multiple-choice – questions or multi-part tables of questions. These surveys, typically fielded to statistically robust numbers of pre-qualified, market-representative panels of respondents, have sample sizes in the hundreds or even thousands.

The validity of quantitative market research surveys, both those aimed at consumers (B2C, or Business-to-Consumer) and those aimed at organizational executives or staff (B2B, or Business-to-Business), rests on the relevance, earnestness and honesty exhibited by survey respondents. In the days when these surveys were fielded over the phone with the aid of live interviewers (Computer-Assisted Telephone Interviewing, or CATI), the interactive human component in the survey-taking process made it relatively easy to spot and eliminate survey respondents that were not relevant, earnest, or honest – those whom we will call "Deceptive Respondents" or **Deceptives**. However, the increasing dominance of online-screened and online-fielded surveys in the last two decades, driven by improved economics and efficiency, has made the task of eliminating deceptive respondents more difficult. This task relies on two major pillars:

1. **Thorough online screening** of participants through a battery of close-ended questions and quotas that are designed to ensure the relevance of market-representative participants automatically selected to take the rest of the survey. These screening questions may include "triangulation" questions so as to "catch" survey takers with contradictory or inconsistent replies to different questions, as well as one or more open-ended questions whose replies ensure respondent lucidity, competence and earnestness. Survey length devoted to screening varies widely, but it can be up to 30% of the entire survey.

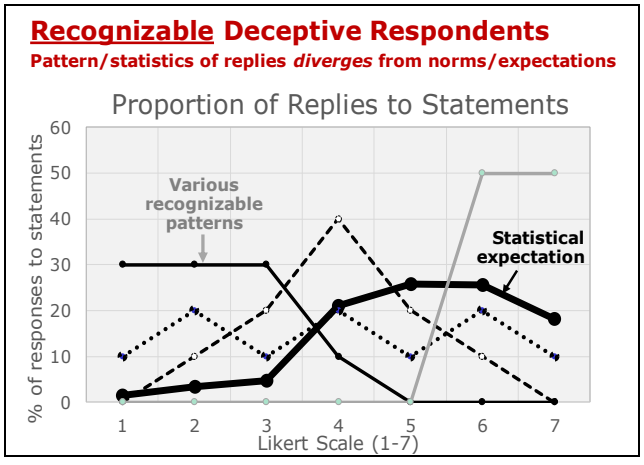
2. **Algorithmic assessment of reply patterns** of each individual respondent to all survey questions, which is aimed at identifying deceptive respondents because of suspicious or implausible characteristics in their replies when compared in aggregate over the whole set of survey questions. For example, scale-based questions (e.g. Likert Scale) may reveal so-called "straight-liners"—those who predominantly give one particular answer (e.g., a "6" or a "1" in a 1-7 scale), those who give patterns of replies to consecutive questions (e.g., 5-6-5-6-5-6-5…) and those who provide non-statistical or random answers. A typical 20-minute online survey can contain upwards of 50 Likert scale type questions or statements, so each completed survey contains sufficient data so as to make these algorithmic assessments.

Both of the above steps, when used thoughtfully in survey design and data clean-up, can be effective at removing potentially deceptive respondents. In particular, assessment algorithms keep improving in their effectiveness over time: the expected statistical distribution pattern of a Likert scale question, for example, can be deduced from the overall survey results – or even from prior survey experiences in a particular market – and respondents whose survey patterns are radically different from a broadly defined expected distribution pattern can be, thus, automatically removed.

Despite its impressive and increasing effectiveness at identifying and removing non-statistical or biased respondents, the algorithmic approach has the limitation of deceptive respondents whose overall response patterns statistically conform – in a broad sense – with expected distributions of responses. Such respondents cannot be thrown out because their response patterns are valid in a statistical sense. However, because their response choices to individual questions are nonsensical or irrelevant, they add only "noise" to the survey results – and they do not contribute statistically to the prioritization patterns or other useful results that one hopes to obtain from the survey. We term these respondents "Intractable Deceptive Respondents" (IDRs).

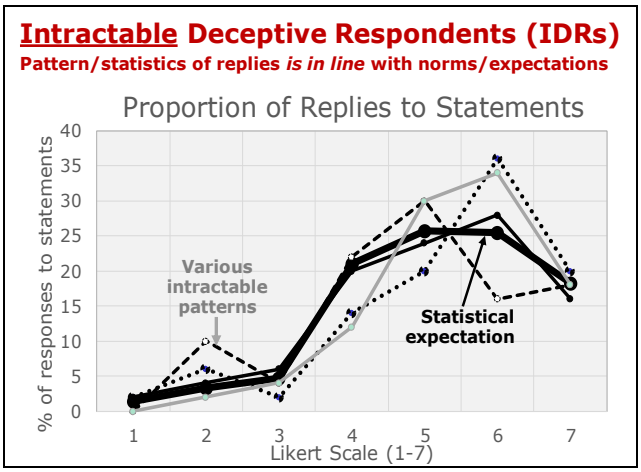
Figures 1 and 2 show the **two types of deceptive respondents**, in a 1-7 Likert scale question example with results typical for a B2B setting in the US market. When assessed across all questions on the survey, response patterns of Recognizable Deceptives, as on Figure 1, are either recognizable in their regularities or far removed from statistical expectations. They can, thus, be removed with the use of ever-improving identification algorithms.

Figure 1: Algorithmically removable Deceptives.



Response patterns of Intractable Deceptives (IDRs), however, as shown on Figure 2, are in line with statistical expectations, even though the data they provide is useless since they just conform randomly with expectations. Identification algorithms do not work for IDRs because they result in too many "false negatives".

Figure 2: Intractable Deceptives (IDRs).



Enhanced or Extreme Screening

The first, and most obvious, approach to managing IDRs is **effective a priori removal through enhanced or extreme screening**. Enhanced or extreme screening relies on excessively long screeners with a high degree of redundancy. This may include several open-ended questions that are reviewed by human "graders"/ assessors, and/or two-stage screening whereby an initial online screener is followed by a telephone interview conducted by specially trained personnel.

If designed and conducted thoughtfully, enhanced or extreme screening can be highly effective in identifying and avoiding IDRs altogether. Very importantly, the process can also boost the research team's confidence that they are indeed including only relevant participants in the survey sample.

Although highly effective, this approach has **two serious drawbacks**:

1. **The length and complexity added to the screening process discourages many legitimate respondents** from participating in the survey, thus lowering Incidence Rates (IRs) and the resultant numbers of completed surveys that can be achieved.
2. **The screening process becomes much costlier and more time consuming**, thus making a portion of survey-based research projects less economical, perhaps even unattractive.

These two drawbacks "feed" on one another, leading to smaller sample sizes. For example, in B2B one may aim for sample sizes of 300 to 500, whereas with extreme screening and its increased time and cost the resulting sample size may be 100 to 200. Smaller sample sizes will

limit the ability to conduct valuable data cuts and extended analysis that is so vital to revealing the deeper insights required in today's complex marketplace. Furthermore, concerns are likely to be raised on the market representativeness of the "extreme screening respondents" due to the incentive amounts required to attract these individuals and the possible profiles and motivations of people (especially among busy IT Decision Makers) willing to undergo the process, even if they are fully relevant to the survey subject matter.

Limited Toleration Approach

To avoid the incremental time, cost and limited sample sizes of extreme screening, it may make sense to explore a second approach, namely the degree to which a certain proportion of IDRs can be tolerated in a survey sample in the sense that their presence in the data will not materially affect the key insights from survey results.

To explore this alternative approach, IntelliClear has undertaken a broad series of **statistical simulations of the effect of IDRs** in survey data, based on years of experience with large scale surveys. The exploration begins with a typical multi-part Likert scale question that is a realistic representation of results that surveys tend to encounter. Our question consists of ten parts, i.e., ten statements that are being tested and need to be cross-compared and prioritized based on the scores they get on a 1-7 scale. We start with the answers that we should expect if the *entire* market had been tested (i.e. no statistical sampling involved).

This **starting point is shown on Figures 3 and 4**. The various percentages on Figure 3 have been selected so as to reflect the variation that we would expect in replies to these ten "competing" statements (A through J), under the assumption that the statements are indeed mildly differentiated from one another – just like we see in real-life survey situations, and in line with cultural norms as to how 1-7 scale survey questions tend to be answered.

Figure 3: 1-7 replies by Statement, adding to 100%.

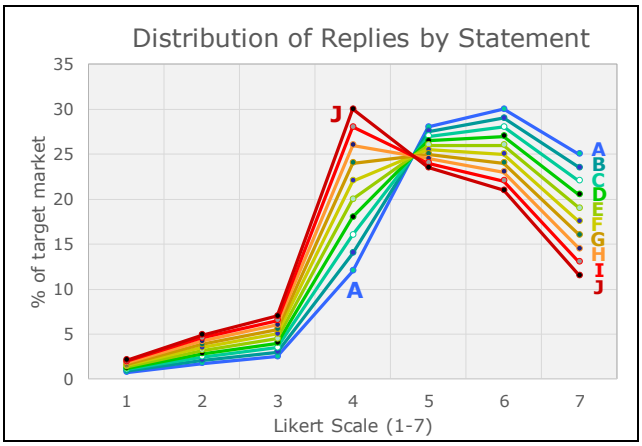
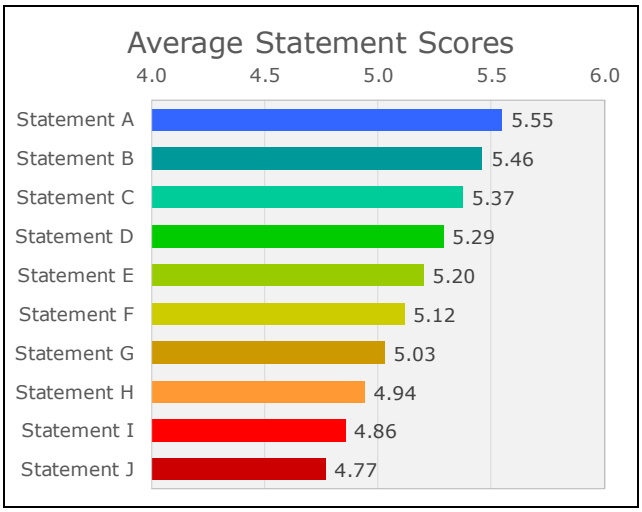


Figure 4: Resultant average for each Statement.



Having defined our market, we simulate it with various sample sizes (Figures 5, 6). Crucially, these simulations assume that all respondents are legitimate respondents that represent the market, thus answering survey questions in earnest. In other words, these simulations do not include any Deceptive participants at all.

Figures 5 and 6 show the aggregate results from a large number (50) of simulations each with two sample sizes: N=300 (Figure 5) and N=600 (Figure 6). As expected, the aggregate averages closely approximate the results for the entire market that were shown on Figure 4. However, the sample size limitations introduce error bars to individual "runs" with those sample sizes, which are also shown, as standard deviations, on Figures 5 and 6.

Figure 5: Average results and error bars, N=300

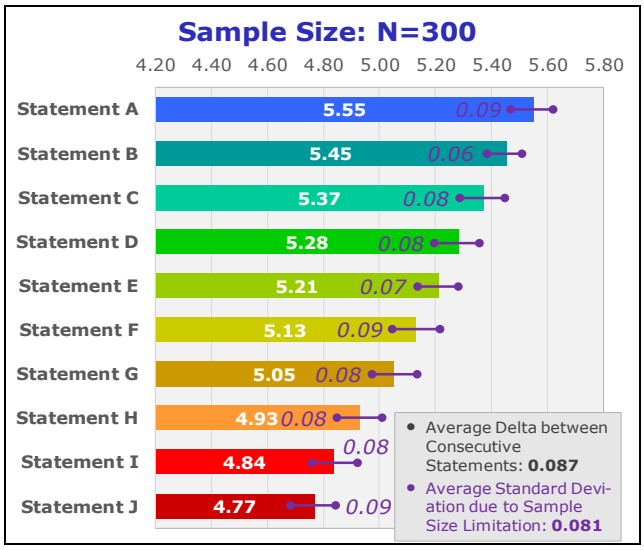
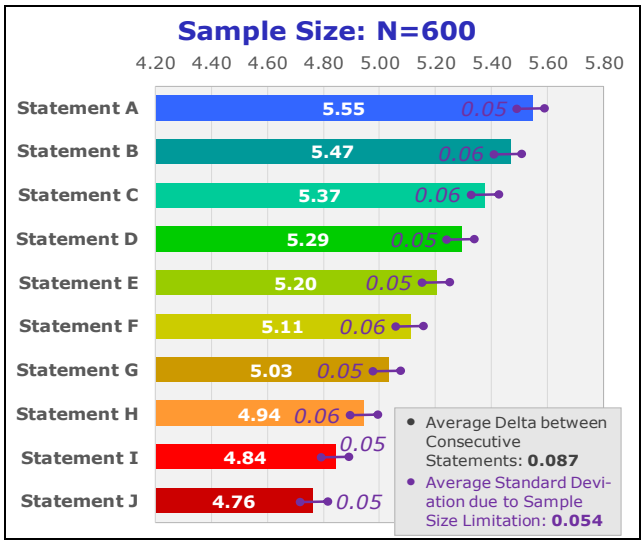


Figure 6: Average results and error bars, N=600



The two preceding Figures point out that to **determine whether a certain sample size is adequate** the crucial factor is that the average delta between consecutive statements (which creates the priority order between statements that the survey is aiming to reveal) should be greater than the average standard deviation associated with the sample size used. As long as the error bar stays below the delta – as is the case on both Figures 5 and 6 – we can reasonably expect that statistics, once the survey is fielded, will not seriously upend the statement prioritization order that we are aiming to reveal.

More specifically, the Figure 5 (N=300) error bar implies that, statistically, the resultant priority order of the ten statements may be upended on 1.5 to 2 occasions on average (out of the 10 statements), but always with modest effects (i.e., by one or two positions); in other words, the general priority order is respected and will be revealed. With a sample size of N=600 (Figure 6) the number of minor upends drops to about 1 occasion on average. Specific examples of these upends, using two consecutive random simulations of survey runs, are shown on Figures 7 and 8.

Figure 7: Two consecutive simulations, N=300

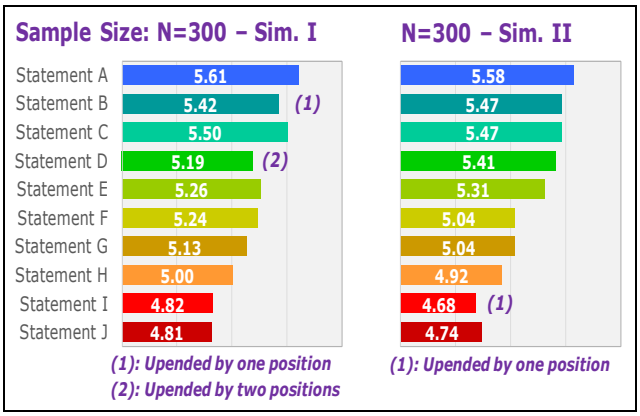
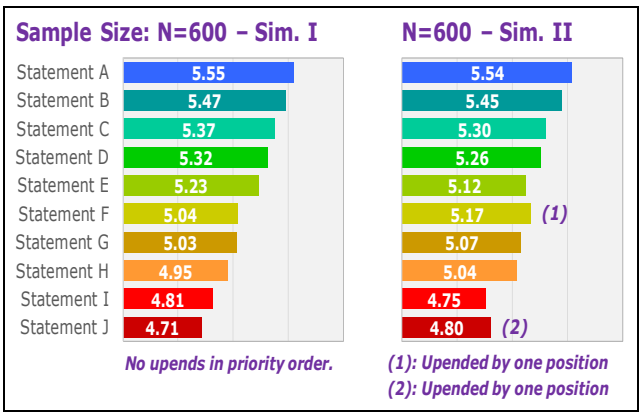


Figure 8: Two consecutive simulations, N=600



Simulation of Deceptive Respondents

As a next step, we can now simulate the **introduction of various proportions of deceptive participants** so as to determine the impact to survey results. Since "Recognizable" deceptives can be readily eliminated algorithmically, we will focus our attention on adding Intractable Deceptive Respondents (IDRs), as discussed. Adding IDRs **dilutes** the distinctiveness of the various statements, i.e., reduces the average deltas between consecutive statements. The addition can also reduce statistical error due to the resultant sample size increase, but far more gradually.

The results are shown on Figures 9 and 10: we start with N=300 and N=600 "legitimate" respondents (0% IDRs, shown at the left side of each chart) and add increasing proportions of IDRs (from left to right).

Figure 9: Proportions of IDRs Added to N=300

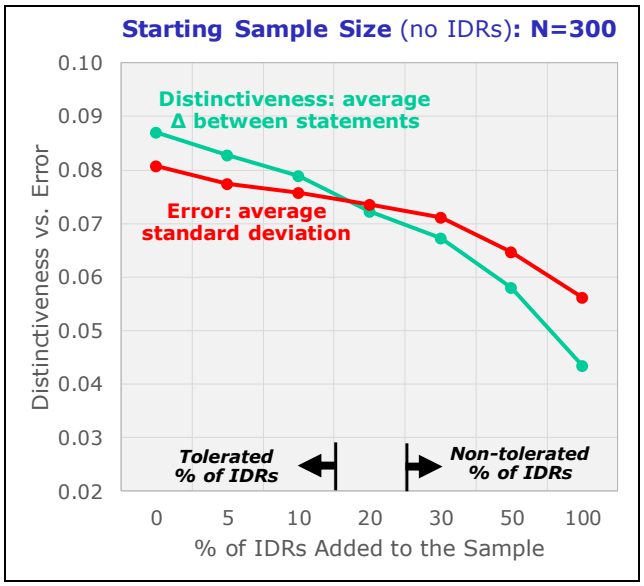
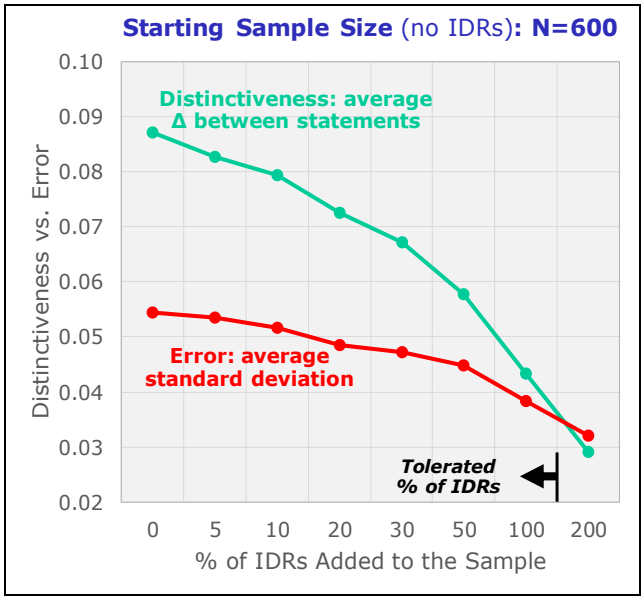


Figure 10: Proportions of IDRs Added to N=600



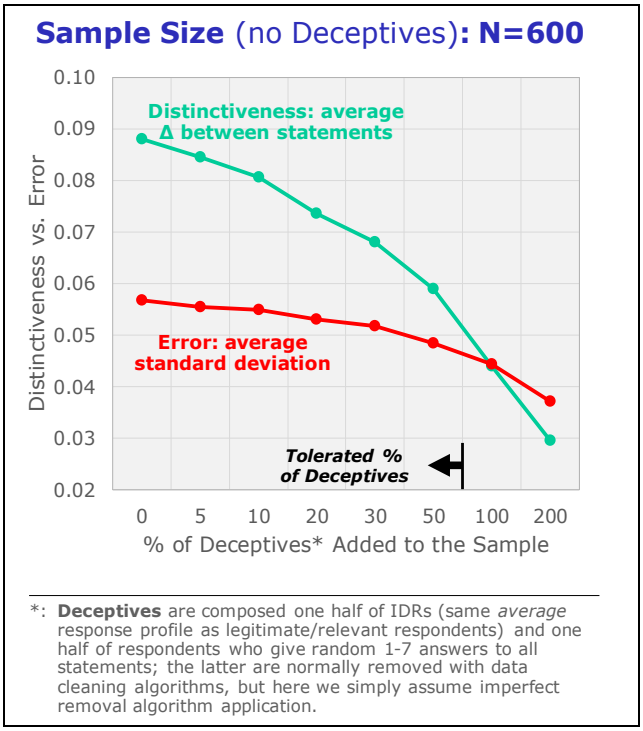
The key conclusions from those IDR-inclusive simulations are the following:

1. Inclusion of a certain proportion of IDRs to the sample does not materially impact results for a surprisingly broad range of IDRs

2. Doubling of the sample size, even though it reduces error bars by about a third, increases dramatically (tenfold in this example) the proportion of IDRs that can be tolerated

An obvious question is: How well do these results and tolerance levels hold if the deceptive respondents are not just IDRs, but a mixture of IDRs with statistically recognizable deceptive respondents (who were somehow not algorithmically removed)? This scenario is simulated in Figure 11, where we see that the relatively high levels of toleration of Deceptives continues to hold true, even though the statistical error bars are by necessity somewhat higher.

Figure 11: Adding various Deceptives to N=600



*: **Deceptives** are composed one half of IDRs (same *average* response profile as legitimate/relevant respondents) and one half of respondents who give random 1-7 answers to all statements; the latter are normally removed with data cleaning algorithms, but here we simply assume imperfect removal algorithm application.

Figure 11 effectively shows that if we have 600 "real" respondents, then in a total sample of 900 survey participants one third of Deceptives (i.e., 300) can be tolerated without effectively impacting results in terms of the relative prioritization of the ten statements. If Deceptives are overwhelmingly IDRs (e.g., Figure 10), the proportion that can be tolerated rises to almost half.

These findings demonstrate the **viability of an online survey data collection strategy that tolerates non-trivial proportions of Deceptive participants**. A rule of thumb implied might be that one should aim, conservatively, for a ~50% greater sample size than planned so as to be reasonably certain that Deceptives can be tolerated in difficult-to-recruit surveys. The point here is not to skip the diligence necessary to remove Deceptive participants through relevant screening processes and identification algorithms. Rather, the focus should be that if such algorithms fall short, and the cost, time and Incidence Rate implications of extensive/ exhaustive two-stage screening are not favorable, then toleration of Deceptives – and, especially, IDRs – is viable, provided the total sample size is reasonably robust.

Ensuring Valid IDR Proportions in Surveys

The primary challenge in deploying a toleration strategy is ensuring **the IDR proportion in the survey sample is not greater than what can be tolerated** without materially impacting survey results.

The difficulty in addressing this challenge lies in the fact that IDRs, because of their statistical profiles, cannot be distinguished from legitimate survey respondents. Therefore, we cannot just identify them and throw them out without risking throwing out truly legitimate survey takers in the process of doing so. What we can only do is identify – conservatively – who *might* be an IDR based on their overall response patterns to the survey, and simply ensure that the resultant numbers of *potential* IDRs are not greater than the limits discussed here.

To identify the potential IDRs, two steps can be followed after the survey results have been collected:

- **STEP 1:** Ensure that all multi-part survey questions (e.g., the questions that seek prioritizations among multiple statements) all lead to reasonably clear – statistically speaking – lists of prioritized statements. This can be done by prioritizing the statements in each question using the entire sample, and then making sure that the same priorities largely emerge when the sample is split into two statistically equivalent pieces. In case two statistically equivalent sub-sets consistently lead to highly diverging statement priorities across most or all of the multi-part questions, then survey results cannot be trusted. However, if the prioritized lists "hold" in most or all cases, then this shows that the survey results can likely be trusted, and one can proceed to STEP 2.

- **STEP 2:** Prioritize statements to all multi-part questions from every survey taker's responses, and identify survey takers for whom the priority lists for all multi-part questions are completely different from the overall priorities ascertained in STEP 1. Even though these survey takers cannot be dismissed, their number likely represents a reasonable upper estimate of potential IDRs. If that proportion is not substantially higher than the proportions discussed in this White Paper, then the potential IDRs can clearly be tolerated with no problem.

This discussion, and the process outlined here, represents a reasonable and viable alternative to extensive or exhaustive screening measures, leading to more natural, relaxed – and, thus, more representative and less biased – survey taking experiences in B2B and B2C.

About IntelliClear

IntelliClear specializes in commercial IT and consumer electronics markets, with an emphasis on the small and medium business (SMB) and large enterprise markets. With a stellar track record of developing results-oriented market segmentation strategies, IntelliClear utilizes powerful data synthesis and seasoned IT industry experience to deliver unique real-world solutions to even the most complex business problems. Through our experienced global partner network, IntelliClear can extend its services into over 65 countries across the globe including North America, Western and Eastern Europe, Asia Pacific, and Latin America.