# ESOMAR/GRBN GUIDELINE FOR ONLINE SAMPLE QUALITY

CONSULTATION DRAFT April 2014

ESOMAR
WORLD RESEARCH

GLOBAL RESEARCH
BUSINESS NETWORK
APRC • EFAMRO • ARIA

ESOMAR, the World Association for Social, Opinion and Market Research, is the essential organisation for encouraging, advancing and elevating market research.

GRBN, the Global Research Business Network, connects 38 national research associations and their over 3500 company members on five continents in promoting and advancing the business of research.

# ESOMAR/GRBN GUIDELINE FOR ONLINE SAMPLE QUALITY DRAFT

## CONTENTS

## 1 INTRODUCTION AND SCOPE

With the emergence and general acceptance of online sources, including panel and river sources, as primary sources of sample for market, opinion, and social research comes a responsibility to measure and ensure the quality of research results using such samples. A number of concerns about sample quality have been raised across the industry. They include concerns about:

- Professional respondents who through various means try to maximize their survey opportunities;
- inattentive if not deliberately untruthful respondents;
- the potential for duplicate respondents as online sample providers broaden their sources in search of more diverse samples and low incidence populations; and
- representativeness, that is, the degree to which research results accurately reflect the target population however defined.

This Guideline responds to those concerns by providing guidance on the operational requirements for the provision of online samples for market, opinion, and social research. It sets out methods to be used by online sample providers, buyers, and end clients to ensure that a sample meets widely accepted quality criteria. It is recommended reading for all stakeholders in the research process, from survey designers to data users.

It is meant to apply to all types of online samples, including those recruited from panels, from social media, and by web intercept methods. It is not meant to cover client-supplied samples such as lists of customers. Nonetheless, researchers may find some of its suggested practices useful when working with these sample sources.

It draws on a number of sources in its basic principles including the CASRO Code of Standards and Ethics; the ICC/ESOMAR Code; the ESOMAR 28 Questions to Help Buyers of Online Samples; ISO 20252 – Market, opinion, and social research; and ISO 26362 – Access panels in market, opinion, and social research. Accordingly, this Guideline recommends full transparency with clients. It encourages the industry use of a common set of terms and definitions. It offers a set of suggested practices for online sample quality, although it is not meant as a substitute for either ISO 20252 or ISO 26362.

This Guideline is not intended to be inclusive of all factors that might impact online research quality. Its scope has been limited to the following areas that are integral to online sample quality:

- Respondent validation
- Survey fraud prevention
- Survey engagement
- Category exclusion (within and across sample sources)
- Sampling (including sample selection, sample blending, sample mixing, weighting, survey routers, sample and panel management, profiling and screening).

Throughout this document the word ‘must’ is used to delineate key principles that researchers are required to incorporate into their work in order to comply with the key legal and ethical practices that define self-regulation in the practice of market research. The word ‘should’ is used when describing recommended practices that operationalize those principles, although other practices may also satisfy the legal and ethical requirements.

## 2 DEFINITIONS

**Access panel**

Database of potential respondents who declare that they will cooperate with future data collection requests if selected

**Completion rate**

The number of respondents who fully complete a survey divided by the number of respondents who begin the survey or at least view the first page

**Cookies**

Cookies are text files containing small amounts of information, which are downloaded to a computer, mobile device or other device when a user visits a website. Cookies are then sent back to the originating website on each subsequent visit, or to another website that recognizes that cookie.

Cookies are useful because they allow a website to recognize a user's device and make website navigation more efficient, remembering user preferences, and generally improving the user experience. They can also help to ensure that offers a user gets online are more relevant to them and their interests.

**De-duplication**

For access panels, a process to remove individuals who are registered more than once on the same access panel so that they are entered only once

For survey samples, a process to remove individuals who complete, or attempt to complete, the same survey more than once

**Device ID**

A technology-enabled system that establishes a set of configuration data about a respondent's device (computer, smartphone, etc.), which can be used to create a machine or device fingerprint. Such systems assume the "machine fingerprint" uniquely identifies a device using settings and characteristics associated with an individual device or, potentially, an individual user account. Device ID systems apply to computers, mobile devices, and other devices that may be used to access the Internet where surveys can be completed.

Note: Device ID is also referred to as "digital fingerprint," a "machine fingerprint" or a "machine ID"

**Duplication**

A situation where a respondent attempts to complete, or completes, a given survey more than once. This can occur, for example, when a panelist or respondent is a member of more than one panel or sample source (panel or sample source overlap) and is selected to participate in a survey that is split across sample sources and fails to recall previously participating in a given survey.

**Consent**

Agreement by a respondent or participant to participate in research, made with complete knowledge of all relevant facts--the information to be collected, how it will be used, with whom it will be shared, and in what form it will be shared. A participant may withdraw consent at any time. The agreement to participate can be collected in written or electronic form. A record of the agreement and how it was obtained must be kept.

**Online sample provider**

A service provider responsible for the provision and management of online samples from relevant sources including panels, web intercept based sources (including river sample sources), email lists, etc. An access panel provider is considered to be an online sample provider.

**Participation rate**

Number of panel members who have provided a usable response divided by the total number of initial personal invitations requesting members to participate in the case of an access panel that relies exclusively on such invitations. Defining participation rate for "river" and other, non-email based approaches is more complicated, with no approach as yet identified as a best practice.

Note: A usable response is one where the respondent has provided answers to all the questions required by the survey design. Where it is possible to determine undelivered invitations (e.g. returned to sender owing to a full email inbox, incorrect postal or email address, or invalid phone number), then these should be taken into account when calculating the participation rate. The number of panel members who did not receive an invitation would then be subtracted from the total number of panel members invited to participate.

**Passive validation methods**

Internet-based methods used to measure respondent characteristics for respondent and panelist validation. These methods may include tracking respondent and panelist website visits, the specific pages they visit, and the links they click, then using that information to create a profile. This profile then is used to provide validation information for respondents and panelists.

**Router**

An online software system that screens incoming respondents and then uses those screening results to assign respondents to one of multiple available surveys. A router can also be used to offer respondents additional screeners and surveys after a screener qualification failure or survey completion. Routers are generally defined as *serial* or *parallel*.

A *serial* router generally utilizes a process in which a respondent is screened sequentially for the available studies on the router. If a respondent qualifies for a survey, the respondent is immediately sent into that survey. If a respondent fails to qualify in a screener, another screener is served to the respondent – with the process repeating until qualification occurs. As stated above, a respondent may also be offered additional screeners and surveys after a screener qualification failure or a survey completion.

A *parallel* router generally utilizes a process in which a respondent is initially exposed to a set of pre-defined profiling or refinement questions and/or a set of pre-screening questions derived from all or a subset of the surveys on the router. After the respondent answers these questions, he or she is assigned to one of the surveys for which he or she appears to be pre-qualified. As it is may be the case that the survey to which the respondent is assigned has additional screening criteria, it is possible for the respondent to fail to qualify for that survey. In that case and as stated above, a respondent may also be offered additional screeners and surveys after a screener qualification failure or a survey completion.

**Representativeness**

The degree to which a sample reflects the target population being studied. A representative sample is one in which the distribution of important characteristics is approximately the same as in the target population. The definition of "important characteristics" generally is a function of the survey topic(s).

**River sample**

A sample where survey respondents are invited via the placement of ads, offers or invitations online

Note: River sample may also be referred to web intercept, real time sampling, and dynamically sourced sampling.

**Sample**

Subset of the target population from which data are to be collected

**Sample broker**

A provider that purchases and resells sample, often providing additional services.

NOTE: A "sample broker" may also be referred to as a "sample aggregator" if they combine or aggregate multiple sample sources.

**Sampling frame**

A list of population elements or other appropriate sources from which a sample is to be drawn

**Sample blending**

The practice of combining multiple, heterogeneous sample sources with the aim of achieving a more consistent or more representative sample. This practice typically utilizes balancing techniques at sample selection and may utilize sample profiling, scoring, or matching techniques.

**Satisficing**

A survey taking behavior in which the respondent gives less than full cognitive effort when answering survey questions. Example behaviors include choosing no-opinion response options; choosing socially desirable responses; straight-lining in matrix questions; and acquiescence, that is, the tendency to agree with any assertion regardless of content.

# 3 KEY REQUIREMENTS

## 3.1 The claimed identity of each respondent should be validated

Researchers and clients have long shared the concern that the incentivized nature of panels may encourage some people to claim numerous false identities as a way to maximize their survey opportunities and subsequent rewards or incentives. Therefore, online sample providers should validate the claimed identify of every respondent. Access panels should be validated at the registration stage and periodically during the stage of individual surveys. Where possible, river samples should be validated at the individual survey stage. Sample brokers should require that their suppliers indicate which sample members have been validated and which have not.

The specific variables used in validation may vary depending on the validation method used, the sources available for validation, and the restrictions, if any, imposed by local laws and regulations. Frequently used variables include:

- Full name
- Postal address
- Telephone number
- Date of Birth
- Email address

If the variables listed above are not available or applicable law and regulation prohibits their use, the online sample provider may use other appropriate methods, including passive methods (provided that such passive methods are not prohibited by applicable law and regulation).

This guideline recognizes that samples drawn from non-panel sources such as river sampling and some forms of routing pose significant validation challenges, as do those drawn in many countries where some forms of external validation are prohibited by law. While passive methods may be possible, their effectiveness has yet to be

demonstrated. In all cases, the specific sources and methods used, any difficulties encountered, and validation outcomes must be documented and shared with clients upon request.

An expanded set of variables may be required when dealing with specialized populations such as physicians or other professionals. This expanded list may include but is not be limited to:

- Full name
- Business postal address
- Business telephone number
- Business email
- Appropriate available professional identification numbers (if relevant and available)
- Professional specialty (if relevant and available)

The data sources used for validation also may vary based on such factors as the type of target respondent or the geographic area being studied. This guideline recognizes that data sources available and useful in one country may not be available and useful in other countries. It also recognizes that the data sources used for validation are seldom all inclusive and that techniques for automatically matching identifying information collected from survey respondents with such sources can result in false positives as well as false negatives. Therefore, multiple data sources should be used where they exist. Further, given these inherent uncertainties in validation outcomes, online sample providers are encouraged to develop outcome codes that express the level of certainty of the identity of each participant (rather than the use of a simple binary indicator indicating success or failure).

The specific sources used for validation must be documented and provided to clients upon request.

### 3.2 Providers must make every effort to ensure that no respondent completes the same survey more than once

As sample providers increasingly use multiple sources (multiple panels, social networks, river samples, etc.) to develop their samples it becomes increasingly likely that the same respondent(s) may be invited to and possibly complete the same survey more than once. Duplicate respondents must be removed prior to analysis, either by the sample provider or the researcher.

#### 3.2.1. Device ID

One common method of de-duplication uses the Device ID from a respondent's computer or device. Device ID is often referred to as a "digital fingerprint," a "machine fingerprint" or a "machine ID." The Device ID is typically created using variables or parameters from a web browser and typically includes:

- Operating system
- Time zone
- Language
- Browser type
- Browser parameters
- Flash ID
- Cookie ID
- IP address

In addition, the Device ID technology used should be capable of supporting geo-location identification and both duplicate and proxy server identification where possible.

The use of Device ID has raised privacy concerns in some jurisdictions. Online sample providers and researchers must ensure that any use of the technology complies with local laws. The Device ID technology used must not access any personally identifiable information and only the Device ID itself can be transferred or stored in a database.

Unfortunately, de-duplication methods that rely on Device ID can be problematic. As with validation of identity, both false positives and false negatives are possible. The increased use of mobile devices to complete surveys means that a more limited set of browser parameters are available to construct a Device ID. As a result, online sample providers and researchers are encouraged to develop outcome codes that express the level of certainty that two or more respondents are duplicates (rather than the use of a simple binary indicator).

### 3.2.2. Cookies and other similar objects

Online sample providers routinely use or cooperate with third parties that use cookies and other similar objects, including local shared objects (commonly referred to as “flash cookies”), web beacons (including transparent or clear gifs) for panels and surveys. The legitimate use of cookies and other similar objects include:

- Identification of panelists or respondents that are required for services requested by the user (to participate in panels and surveys)
- Validation and fraud prevention, including legitimate use in Device ID technologies
- Advertising evaluation and tracking research

When cookies and other similar objects are used in panels and surveys, online sample providers and researchers must comply with all applicable laws, regulations, and industry codes, including the separation of research and marketing activities. In some jurisdictions, this includes obtaining panel member and respondent permission to place cookies and other similar objects on their devices for the first time. Respondents must be told what cookies and other similar objects are and why they are being used. This information must be presented in language that is easily understood so that panelists and respondents can make an informed choice about whether to give their permission.

### 3.2.3. Other Methods

Online sample providers may use alternatives to Device ID technology, cookies, and other similar objects if they accomplish the equivalent functions at the same level of accuracy and effectiveness. These methods include other technology solutions as well as process-based solutions. Any alternative method must of course comply with local laws and regulations.

Regardless, the method used must be fully documented, and the results of the de-duplication process provided to the client upon request.

## 3.3. Respondent engagement should be measured and reported on.

There is widespread concern among clients that online surveys are especially vulnerable to questionable data supplied by respondents who do not give adequate level of thought to answering survey questions or deliberately provide fraudulent answers. It is important to identify these respondents so that their impact on a study’s overall findings is minimized.

Research on research has identified a broad set of possible measures that may be used to identify inattentive respondents. These include, but are not limited to:

- Survey completion time
- Proportion of unanswered questions
- Extent of selection of non-substantive answers such as “Don’t Know” or “Refused”
- Patterned responses in matrix or grid questions (e.g., straight lining, random responding, etc.)
- Detection of inconsistent responses such as asking both positively and negatively worded questions in the same attribute battery
- “Red herring” questions such as “Check the box on the far right”
- Appropriate responses to open-ended questions

The researcher designing the survey and the company hosting it generally share responsibility for identifying potential inattentive or fraudulent respondents. The appropriate division of responsibilities is a matter to be negotiated between the two parties. The use of multiple measures from the above list is strongly recommended along with a scoring system that aggregates these measures across each respondent with the scores being included on the respondent data record. Researchers and clients should work together to determine the specific measures to be used as well as the threshold in survey scores that determines which respondents, if any, are deleted. The online sample provider should be prepared to replace any respondents whose data is deemed unacceptable by the client.

The measures used and the method of calculation of the overall score must be documented and shared with the client on request.

### 3.4. The identity and personal data of respondents must be protected.

Any and all data collected from respondents online must be kept securely and only used for market research purposes. No personally identifiable data may be shared with a client without the consent of the respondent, and when done so it must be in compliance with local laws, regulations, and industry codes. When consent is given to share data with a client, the responsibility to keep the data secure and protect the identity of respondents transfers to the client.

### 3.5. Take special care with children and young people.

Online sample providers must ensure that no child is selected for a research project unless a parent or legal guardian, or other person legally responsible for the child has given permission for that child to participate in the specific project for which he or she is sampled. The legal definition of a child varies substantially from jurisdiction to jurisdiction and the sample provider must comply with the laws in the jurisdiction in which the child lives. Where there is no specific national definition, those aged under 14 should be treated as “children” and those aged 14-17 as “young people.” These age ranges generally recognise the different stages of mental and psychological development.

Buyers of online sample must take care to ensure that the appropriate permissions have been obtained prior to interviewing children.

### 3.6. Employ good online questionnaire design practices.

Despite almost two decades of online research and a significant body of research on research about online questionnaire design there are few widely accepted best practices. For example:

- The longer the questionnaire, the more likely respondents will disengage and potentially jeopardize data quality. A number of studies have shown an increase in satisficing behaviors and even breakoffs after 18-20 minutes.

- Research also has shown that a repeated series of matrix or grid-style questions can result in straight-lining or other patterned responding.
- A phenomenon known as primacy, where questions with a large number of answer categories can result in respondents choosing responses from the top of the list more often than the bottom.

In general, the best questionnaire practices are those that result in interesting and easily understood well-designed questions with an equally well-designed list of potential answer choices presented in a logical order.

### 3.7. Online sample providers must be transparent with researchers and clients about sample sources, the sampling process, and its outcomes

If users of online samples are to have confidence that their sample is fit for purpose, then online sample providers must make available information about the sample development process. This includes:

- A description of the sampling frame or sources from which the sample was drawn (including any subcontractors used), how it was constructed or acquired, and the target population it is meant to represent;
- the sampling method used to select potential respondents from the sampling frame or equivalent and the means employed to ensure that the sample represents the target population, including any quotas or sample blending methods used;
- the specific criteria used in sample selection, such as quotas or other filtering criteria;
- the incentive offered to sample members;
- where panels are used, a count of the number of sample units drawn and solicited, the number of bounced emails, the number of partial interviews, and the number of full, completed interviews; and
- where a router or similar intercept method is used, a count of the number of the number of potential respondents screened, the specific criteria used and the number of respondents qualifying. When use of a specific router design is known to produce some bias in respondent selection, it also must be documented.

### 3.8. Researchers using online samples must be transparent with clients.

The appropriate reporting standards for research projects using online samples are similar to those for research projects in general. The following should be provided routinely to clients:

- The sampling frame or equivalent, sources and sampling methods used
- The dates of fieldwork
- The average survey length
- The total number of interviews completed
- Any quotas used or other specifications used in sample selection
- The questionnaire and other relevant data collection documents
- A count of the number of survey respondents whose identify was successfully validated
- A description of any de-duplication methods used and the number of responses deleted as a result
- The measures of respondent engagement used and an account of any respondents removed or replaced because of poor survey behavior
- Exclusion information

- Participation rates[1] (where possible) and methods used to calculate them
- Whether all or part of the project was subcontracted and, if so, to what organizations

In addition, providers and users of online panels have additional reporting responsibilities due to the variety of sampling methods used for online research. While some online panels are recruited using traditional probability-based methods, most are not. Recent innovations such as online routers and advances in dynamic sourcing cast a still wider net across the Internet to solicit volunteers to complete surveys. As a consequence, the vast majority of online samples are convenience samples that lack the necessary statistical properties assumed to be needed to accurately represent the intended target population, thereby creating substantial risk that a study's results may contain significant error.

One common practice has been to impose demographic quotas (primarily age and gender) in sample selection or in post survey adjustment. A number of studies have shown that these adjustments are often insufficient and additional adjustments using attitudinal or behavioural variables correlated with the survey topic are needed to improve accuracy.[2]

ISO 20252—Market, opinion, and social research requires that researchers report to clients "the procedure used to select potential respondents from the sampling frame or equivalent and the means employed to ensure that the sample represents the target population." It further requires that researchers describe their weighting and projection methods and provide an "assessment of how well the sample represents the target population and the associated implications for data quality." Similarly, ISO 26362 – Access panels in market, opinion and social research requires that "the access panel provider shall agree with clients on the design and methods to be used to draw samples from access panels for surveys or other research purposes." It further requires that the sampling methods used shall be reported to clients. ESOMAR and GRBN consider these ISO requirements to be best practices that all researchers should follow.

### 3.9. Passive Data collection

Passive data collection refers to research methods that collect data without the traditional use of survey questions. In many cases, these data will be considered personal data. Sources of passive data include web browsing data, app data, loyalty data, geo-location data, social media data and data generated by/obtained from mobile devices. Much of this data can be combined with survey data.

All passive data collection methods utilized must comply with local laws and regulations. As with personal data, researchers in many jurisdictions will have to set out a clear legal basis for using and processing this data, including its use for sampling activities and obtaining the consent of the individuals concerned.

### 3.10. Conform to all relevant laws, regulations, and industry codes of conduct.

It is critical that both online sample providers and buyers be aware of and strictly adhere to all relevant regional, national, and local laws and regulations as well as any industry codes of conduct or cultural dispositions that may set a higher standard than what is legally required.

---

[1] This guideline recognizes that increasing use of river sampling and routers makes calculation of participation rates difficult if not impossible. Until a generally accepted best practice emerges researchers should note this difficulty in their reports to clients.

[2] For further discussion see, "Report of the Task Force on Non-probability Sampling," available on the website of the American Association for Public Opinion Research (www.aarpor.org). A summary of the report along with critiques from several experts in the field has been published in *The Journal of Survey Statistics and Methodology*, Volume 1, Number 2, November 2013.

## 4. PROJECT TEAM

Reg Baker, Co-Chair, Market Strategies International

Peter Milla, Co-Chair, CASRO

Pete Cape, Survey Sampling International

Mike Cooke, GfK

Melanie Courtright, Research Now

George Terhanian, Toluna

## 5. REFERENCES

To be added

**- END OF DOCUMENT -**