

PANEL RESEARCH 2006

ESOMAR WORLD RESEARCH CONFERENCE

BARCELONA / 27 - 29 NOVEMBER



CONFERENCE PAPERS

WORLD
ESOMAR

RESEARCH

ASSESSING INDIVIDUAL RESPONDENTS' QUALITY

AN INNOVATIVE SCORING SYSTEM

Christian Loeb

Adriane Hartmann

WHY DO WE NEED TO MONITOR RESPONDENT QUALITY?

Respondents who fake responses are the nightmare of all panel managers. They enter bogus data in order to receive a participation incentive; they click through questions without reading; they might even try to enter a panel using multiple identities. Some of these cases are noticed ex-post and removed during data cleaning, but some enter the final data and dilute or bias results.

Working with a broad range of recruitment channels like websites or newsletter advertising leads to respondents with different levels of motivation. But even the best recruitment channels may yield a certain number of respondents who try to cheat by either giving random answers or by clicking rapidly through the surveys.

Data cleaning can never remove all cases of cheating. A way of assessing the quality of respondents before they take part in a client's survey would minimise cheating attempts as well as the number of cases of dishonest respondents making it into the sample.

Scoring respondents to increase data quality is not the only part of a comprehensive panel quality management process. It has to be accompanied by other measures such as avoiding the recruitment of respondents from dubious resources such as survey portals like www.getpaidguides.com as well as the prevention of active registrations. (See figure 1.)

This paper explains why a scoring model is suitable for optimizing panel quality. The authors then demonstrate

their approach, and show how scores and variables like motivation and concentration correlate.

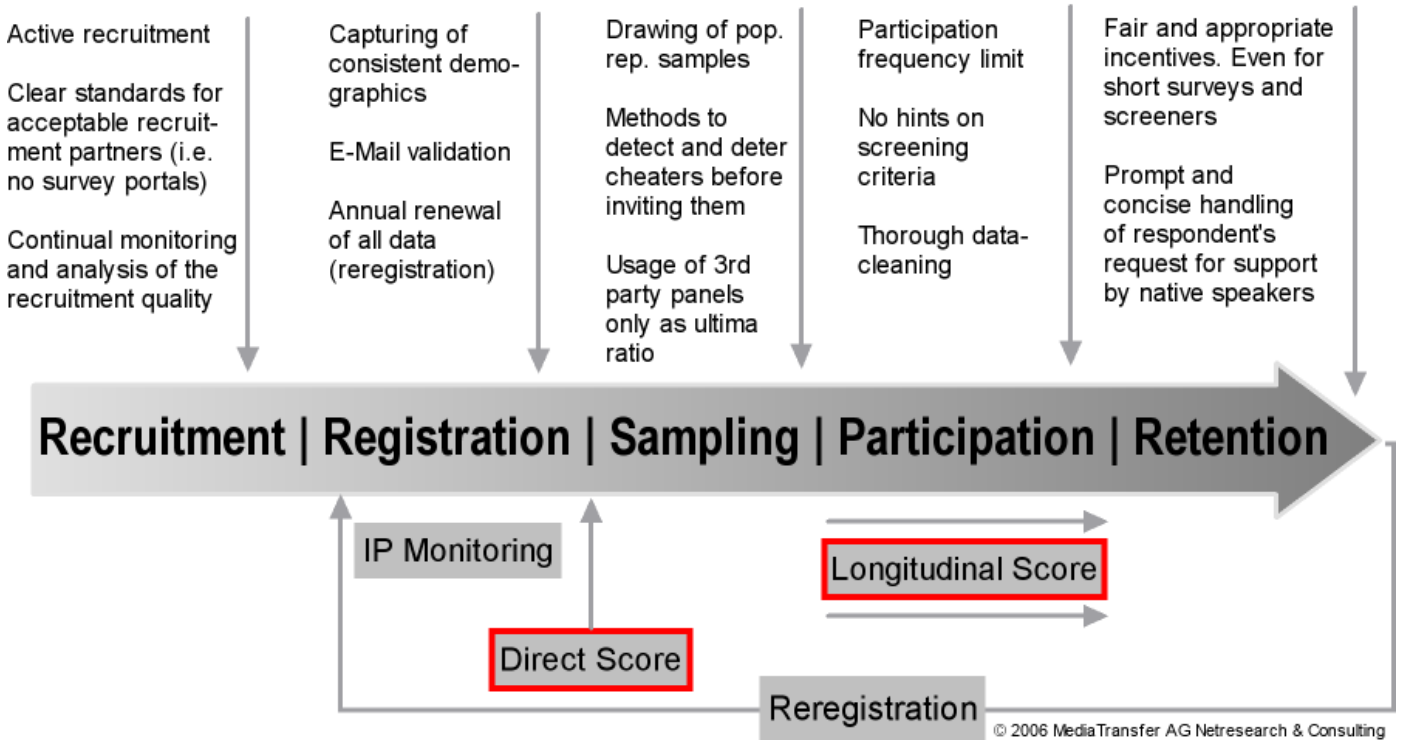
WHY WORK WITH SCORING MODELS?

Our research on motivation for panel participation and the ongoing analysis of our pan-European online panel over the last few years show that there is no single yes/no criterion for detecting whether a respondent behaves as desired and answers truthfully and thoroughly. For example:

- a) Computer savvy users can fill out surveys faster in a more thorough manner than an inexperienced user who is simply clicking through it. Thus, completion time alone is not a suitable criterion.
- b) Every respondent should be allowed to make a typo from time to time. This also happens in the real world outside of surveys and should not be a reason for permanently excluding a panellist.
- c) Respondents' behaviour can often be ambiguous. For example, an otherwise thorough respondent may sometimes start clicking through a survey if the survey is very lengthy.

In light of this, it does not make sense to exclude a panellist from the panel based on any single criterion. For that reason we use a combination of several criteria and aggregate them into a score. Furthermore, a score also allows the weighting of events, e.g., not finishing a survey has a smaller influence on the score than clicking through it. As the variables impacting the score are kept, the weighting of single events and therefore the whole model

FIGURE 1
POSITION OF THE DIRECT AND LONGITUDINAL SCORE WITHIN THE PANEL QUALITY MANAGEMENT PROCESS



can be redesigned and reapplied whenever the model is updated.

In order to assess the respondent's quality before and during his membership we need to have two scores: the direct score and the longitudinal score. The direct score serves as a quality barrier before the first client survey, while the longitudinal score monitors the respondent's quality during membership. We will elaborate on each score in the following.

THE DIRECT SCORE

The direct score process is carried out once during the panelist's lifecycle. Two days after initial registration, each panelist receives an invitation to a survey, a typical concept test. However, this test is not a client survey but instead used to assess the panelist's sincerity. The respondent's individual direct score is calculated based on a comparison to answers given during the initial registration process, comparing answers given within the direct score survey (with a rotated rating-scale),

time to complete, and other variables partially collected on a technical level.

The respective weights of the elements comprising the score are derived from certain normative *a priori* criteria based on the threshold of acceptable response behaviour and have been optimized over time, building on empirical evidence. For example, if the age given in the direct score survey does not match the age given during registration and the respondent gave more than three inconsistent answers, (s)he will not be invited for a second survey. On the other hand if the timing is below a predefined level, but questions are answered consistently, indicating the respondent read and understood the questions, (s)he will become part of the panel.

THE LONGITUDINAL SCORE

Each survey the panelist is invited to impacts the longitudinal score. The score is affected negatively if the panelist is excluded during data cleaning due to inconsistent answers or an implausible completion time.

Positive behaviour like full participation is however taken into account for score calculation as well. The model should not penalise respondents who use the internet rarely. We have therefore chosen to look at the number of full participations as opposed to using an invitation vs. the participation ratio.

In addition, during our annual re-registration the data given by the panelist is compared against the registration of the previous year. Deviations in invariable or typically static demographics impact the score.

VALIDATING THE SCORES - METHOD

The ultimate question is whether this approach increases data quality. Thus, we conducted a survey among panelists with a wide range of scores designed to evaluate this question. A total sample of 1,200 respondents from our panels in the United Kingdom and Germany answered a questionnaire intended to measure consistency of answers on the individual respondent level, motivation, and concentration. Respondents were divided into four groups according to their direct score and into two groups according to their longitudinal score. Each subgroup contained 100 respondents.

Consistency was measured in the following way: At the beginning of the questionnaire the number of persons in the households, the number of pets and the length of the respondent's hair were asked. At the end of the questionnaire, after several other subjects, these questions were asked again this time using slightly different phrasing and answer categories.

Motivation was measured by open-ended questions on brand awareness in three product categories in which numerous brands are usually known (cars, coffee, toothpaste). The number of brands listed was considered to be an indicator of the respondent's motivation.

Concentration was measured as follows: The respondents were exposed to two shampoo packages in relatively quick succession and then asked to name the brand of the first package. After a number of such exercises the instruction was then changed to ask for

the name of the second brand.

The results of this survey are described in the following three paragraphs.

Effect of score on panelists' response consistency

The consistency checks with respect to the hair length question and the pets question delivered only a very small number of inconsistencies across respondents of all score levels. This is presumably due to the rather obvious way in which the questions were asked. The household question, however, yielded some interesting results. Because the second household question was asked in a rather unusual way ("How many people live in your household, yourself EXCLUDED?"), a lot of errors occurred due to insufficient accuracy when answering the question. This type of error clearly was more common in subgroups with bad scores, particularly in Germany (60 % vs. 71 % in the subgroup with the best / worst direct score and 49 % vs. 64 % in the subgroup with the best / worst longitudinal score). Inconsistent answers which could not be attributed to a misunderstanding of the question were quite rare. Although the numbers are low, this type of error is more often found in groups with a bad score, especially in the UK (2 % vs. 4 % in the subgroup with the best / worst direct score, 0 % vs. 3 % in the subgroup with the best / worst longitudinal score).

Effect of score on panelists' motivation

The direct score groups significantly ($p < 0.1$) differed in the number of listed items. The group with the best direct score named 7.1 items on average, whereas the group with the worst direct score named 6.4 items in the car brand question. In the toothpaste question, the relation was 3.7 items to 3.4 items.

The longitudinal score groups differed only in the toothpaste question (4.1 vs. 3.7 items; $p = .01$). Thus, the toothpaste question turned out to be the most sensitive measure of motivation. This is easily explained by the fact that the toothpaste question was the last of three very similar questions and therefore, was rather challenging for the respondents' motivation.

Effect of score on panelists' concentration

In the concentration part of the task, there were 10 trials in which recognition of the first presented brand was requested. The packages were shown in short succession, but the first package was shown long enough (400 ms) to be easily recognized by an attentive and focused respondent. Overall, hit rates were higher for respondents with a good longitudinal score (9.5 of 10) than for respondents with a bad longitudinal score (8.9 of 10; $p < .001$). For the direct score groups, only UK respondents showed a significant relationship (9.3 vs. 8.8; $p < .01$).

In the compliance part of the task, there were five trials in which recognition of the second presented brand was requested. The latter should have been a lot easier as the second package was shown for 1000 ms and because the first package usually generates less interference for the second package than vice versa. However, the change in the task was hidden in the instruction in such a way that only careful and attentive respondents would notice.

It is clear from the hit rates that the tendency to gloss over the instructions and therefore to miss the change is higher in the groups with a bad score than in groups with a good score (direct score Germany: 13 % vs. 18 %; direct score UK 11 % vs. 21 %; longitudinal score Germany 7 % vs. 12 %; longitudinal score UK 10 % vs. 16 %).

CONCLUSION AND NEXT STEPS

Across two countries the survey shows that the two scores differentiate between panelists with respect to consistency, motivation, and concentration. Thus we conclude that exclusion of those with bad scores increases both the reliability and validity of the data. While the absolute effects are in most cases modest, they are nevertheless substantial. Even well-intending respondents sometimes miss instructions or mistype answers, an unavoidable fact furthermore supported by our research. With that in mind, our goal to reduce the additional noise generated by faking respondents

has been achieved. Another goal achieved as a result of using a respondent scoring model is to build a foundation for judging the quality of recruitment partners, allowing us to continuously refine our partner portfolio.

In the future we look forward to incorporating additional variables into the model and of course optimising of the model itself.

At this point in time, the direct and longitudinal scores support our efforts to work exclusively with serious and motivated panelists. As a result, when used as part of a complete panel quality management system, data quality is considerably improved, increasing the value to clients.

The Authors

Christian Loeb is member of the board responsible for panel management, MediaTransfer AG Netresearch & Consulting, Germany.

Adriane Hartmann is a senior consultant responsible for method development, MediaTransfer AG Netresearch & Consulting, Germany.