Global Journal of Quantitative Science
Vol.1. No 1. March 2014.. Issue. Pp 15-29.

Development and Application of a Survey Quality Assessment Model

Tao Chen[1], ***Robert Raeside***[2] and ***Hafiz T.A. Khan***[3]

Abstract

Until recently research has been undertaken to improve the quality and effectiveness of surveys. By reviewing existing works, an improved model has been developed in this paper in order to assess the quality of statistical surveys. The model has been applied to over 103 national and international surveys and it is then used to generate synthetic data based on similar principles such as formulation, design and application quality etc. Various measures of the survey have been undertaken such as average question length, number of questions etc., and input into the database; which are analysed using statistical classification tools. This has allowed the identification of structural features of surveys which is associated with survey quality and this is reported precisely in the paper. Recommendations are made as to how a systematic approach can be taken to improve survey quality and effectiveness.

Key words*: Survey Methods, Analysis, Quality of Survey, Database, Measurement Tools*

1. Introduction

For many decades survey methods have been utilised in almost every research area including in business and social sciences, health sciences and even in engineering. Among these in some areas, especially public opinion, official statistics, and marketing research, survey methods have been becoming more and more popular means of collecting statistical information (Moser and Kalton, 1979; Schwarz *et al.*, 1995; Groves *et al.*, 2009).

With the rapid increase in applications of survey methods especially via e-surveys, the quality of data gathered in the surveys has been in discussion for last two decades. There were some related research projects largely managed by national statistical institutes and led to organisations publishing their own quality guidelines, see for example, Statistics Sweden (Anderson *et al*., 1997), Statistical Policy Office USA (2001), Eurostat (Bergdahl *et al*., 1999), Statistics Austria (Burg, 2004), Statistics Finland (Laiho *et al*., 2002), Rosen and Elvers (1999), Statistics Canada (1998), and Eurostat (2002). In the UK, the Office for National Statistics (ONS) has published its Guidelines for Measuring Statistical Quality (ONS, 2004). These do not represent mandatory procedures for official statisticians, but they provide important guidance for anyone who is involved in producing statistical outputs. In addition, there were six international conferences on quality in official statistics, which took place in Stockholm in 2001, Mainz in 2004, Cardiff in 2006, Rome in 2008, Helsinki in 2010, and Athens in 2012 respectively. More than 2000 delegates attended these conferences, including many from academia and other non-governmental organisations. These demonstrated considerable continuous interest in the concept of survey quality and related research activities.

The guidelines and methods of measuring and assessing survey quality are mainly based on qualitative verbal descriptions and are mainly focused on the survey data quality. They are not easy to implement and follow in practice and it is very difficult to make comparisons between the surveys. However, survey process quality is of more importance because better designed, administered, controlled and implemented processes can definitely lead to great improvement of the validity and reliability of survey data. Due to these reasons there is a demand, theoretically and pragmatically, to design a model which can give a holistic quality assessment and evaluation of not only the data but also the processes of survey. With the intentions of meeting such needs, Chen and Raeside (2008) attempted to build an appropriate model and made contribution towards survey quality evaluation and improvement in theory and practice. Such a model enables researchers to assess the quality of different types of surveys in a survey database, classify surveys in respect of quality, learn from the classification, and make recommendations on how to improve survey quality and effectiveness. Therefore,

[1] *Research Assistant, Employment Research Institute, Edinburgh Napier University, Craiglockhart Campus, Edinburgh EH14 1DJ Scotland, UK, E-mail: T.Chen@napier.ac.uk*

[2] *Professor of Applied Statistics, Employment Research Institute, Edinburgh Napier University, Craiglockhart Campus, Edinburgh EH14 1DJ, Scotland, UK, E-mail: R.Raeside@napier.ac.uk*

[3] *Senior Lecturer in Applied Statistics, Department of Economics and International Development, Middlesex University, London NW4 4BT, England, UK E-mail: H.Khan@mdx.ac.uk. Also Visiting Fellow in Demography, The Oxford Institute of Population Ageing, The University of Oxford, Oxford OX2 6PR, England, UK, E-mail: hafiz.khan@ageing.ox.ac.uk*

the aim of the paper is to develop and calibrate a survey quality assessment model, and apply the model to assess the quality of surveys.

The paper starts with an introduction of background knowledge about survey quality and assessment, then moves on to describe the development, validation and calibration process of the survey quality assessment model and finally presents to use the model to assess the quality of surveys.

2. Survey quality and assessment

The importance of ensuring survey quality

Conducting a survey whether at an international, national to regional level or even a single research project involves every stage of survey process including inception, design, construction, execution, analysis and dissemination. This is a very complex and comprehensive process, usually very costly, needs a long period of time and involves a number of people; therefore ensuring survey quality is of paramount importance. A good quality survey can result in high quality of data which is vital for drawing correct conclusions and is necessary to convince policy-makers. On the other hand, however, a poor quality survey causes invalid and unreliable survey data problems that can bias and distort the survey results and lead to incorrect decisions and misleading information to policy-makers.

Measuring survey quality

The concept of survey quality encompasses several dimensions, most of which have long been considered by survey researchers, but not necessarily in a consistent way and often independently of one another. On one hand, it can be defined as a combination of the representativeness of the sample, the accuracy and precision of measurements, data processing and management with several subcomponents in each (Tolonen *et al*., 2006); on the other, it can be defined as the closeness of the match between a survey's objectives and its outcomes (O'Muircheartaigh, 1997).

Many statistical agencies in the US Statistical Policy Office (2001), took the approach of improving "survey quality" by measuring and minimising various error sources that affect data quality. If the data quality is limited to the accuracy dimension, then different aspects of survey accuracy can be related to various error sources in survey. Five of these error sources can be identified: sampling error, nonresponse error, coverage error, measurement error, and processing error (Osborne, 1942; Dalenius, 1977; Andersen *et al.*, 1979; Assael and Keon, 1982; Bound *et al.*, 2001; Groves, 2004; Weisberg, 2005; Alwin, 2007). Measuring the quality of survey takes on different meanings depending on the constituency. Different survey data users have different goals and, consequently, different ideas of what constitute quality. Similarly, the reporting of "quality" can be implemented quite differently depending on the type of data product produced.

According to Statistical Policy Office USA (2001), one can break the survey quality process into components or characteristics that focus around several key concepts: accuracy, relevance, timeliness, and accessibility. In Eurostat (2002), quality of statistics is defined with reference to seven criteria: *accuracy, relevance, timeliness and punctuality, accessibility and clarity, comparability, coherence and completeness*, which are very similar to the American system and only adding in punctuality and clarity. These characteristics are defined as:

- *Accuracy* in the general statistical sense denotes the closeness of computations or estimates to the exact or true values (Marriott and Kendall, 1990).
- *Relevance* is the degree to which statistics meet current and potential users' needs (Statistics Canada, 1998).
- *Timeliness* According to Anderson *et al*., (1997) this is the length of time between its availability and the event or phenomenon it describes. Timeliness can refer to several concepts including the length of the data collection's production time, the time from data collection until the first availability of a product and the frequency of the data collection.. Timeliness can, however, be difficult to characterise since the characteristics of the data collection can often reflect the availability of data.
- *Accessibility* refers to the physical conditions in which users can obtain data: where to go, who to contact, availability on a web site, format of data (degree of aggregation), costing policy, marketing conditions (copyright, etc.), availability of micro or macro data in various formats (paper, files, CD-ROM, Internet…), etc (Statistics Canada, 1998). Accessibility also implies the data products include adequate documentation and discussion to allow proper interpretation of the survey results, this can be enhanced by provision of user workshops and training. Data products tend to have higher value if they are easily accessible.
- *Comparability* is the extent to which differences between statistics from different geographical areas, non-geographical domains, or over time, can be attributed to differences between the true values of the statistics

(Arondel and Depoutot, 1998). The sources of distortion of comparability arise from two sources, the use of different concepts/definitions and the use of different measuring tools or procedures.

- *Coherence* of statistics is their adequacy to be reliably combined in different ways and for various uses. When originating from a single source, statistics are normally coherent in the sense that elementary results derived from the survey can be reliably combined in numerous ways to produce new variables. When originating from different sources, statistics may not be completely coherent because they may be based on different approaches, classifications and methodological standards (see Arondel and Depoutot, 1998).
- *Completeness* is the extent to which all statistics that are needed are available (Arondel and Depoutot, 1998). It is usually described as a measure of the amount of data made available compared to the amount that was expected.

In Eurostat (2002; 2003), quality of statistics is defined with reference to seven criteria: accuracy, relevance, timeliness and punctuality, accessibility and clarity, comparability, coherence and completeness, which are very similar to the American system and only introduced two new concepts: punctuality and clarity.

- *Punctuality* is defined by Eurostat (20030 is the time lag existing between the actual delivery date of data and the target date when it should have been delivered.
- *Clarity* refers to the data's information environment whether data are accompanied with appropriate metadata, illustrations such as graphs and maps, whether information on their quality also available (including limitation in use) and the extent to which additional assistance is provided by the data provider (Eurostat, 2003).

The Organisation of Economic Cooperation and Development (OECD) (2003) also views its quality of statistics in terms of seven dimensions: relevance, accuracy, credibility, timeliness, accessibility, interpretability and coherence. In which it is introduced another two new concepts, those of credibility and interpretability.

- *Credibility* refers to the confidence that users place in those products based simply on their image of the data producer. Confidence by users is built over time. One important aspect is trust in the objectivity of the data which implies that the data are perceived to be produced professionally and reliability in accordance with appropriate statistical standards, and that policies and practices are transparent. Clearly to have credibility, data cannot have been manipulated, nor their release timed or formed in response to political pressure.
- *Interpretability* reflects the ease with which the user may understand and properly use and analyse the data. This requires adequate definitions of concepts, target populations, variables and terminology, underlying the data, and information describing the limitations of the data, if any, largely determines the degree of interpretability.

Survey data quality is a concept with many dimensions and each dimension is linked with others. In the abstract, all dimensions of data quality are very important, but in practice, it is usually impossible or infeasible to place equally high importance on all dimensions. Thus, with fixed time, labour, access and financial resources, an emphasis on one dimension may well result in less emphasis in another (Groves, 2004); and ensuring survey quality is to find an appropriate method of balancing the entire elements associate with it.

The survey process

In order to develop a holistic audit tool assessing the quality of survey, the survey process is first broken down into eight building blocks comprising of formulation, sample frame, instrument, administration, data entry/validity, quality assurance, report and dissemination. These are displayed in an ordered process in Figure 1. For detailed definition and explanation of these processes please see Chen (2011).

Developing the assessment model

With the intention of holistically measuring the quality of a survey, an assessment model is developed from the hierarch processes presented in Figure 1. The model is essentially an Excel based audit following through all stages of the survey from design, construction, execution, analysis to dissemination. At each stage a set of criteria is applied to make an assessment of quality. Scores attained against these assessments are weighted by the importance of the criteria and summed up to give an overall assessment of the stage. These stage assessment scores are weighted again by the importance of the stages and summed up; therefore the total quality score of a survey can be obtained.

As defined earlier the survey quality consists of eight dimensions, which are not all of the same importance in the survey measuring process. In order to weight the level of importance of these dimensions in respect to survey quality, Delphi (Dalkey, 1969) and Multiple Criteria Decision Making (MCDM) (Belton and Stewart, 2002; Saaty, 1985) methods were applied obtain the consensus opinions on the weights.

The method used here is to assign weights to assess surveys of different types applies to a specific MCDM procedure called value tree analysis. The decision-making framework is assumed to be a hierarchical weighting model. First objectives are structured hierarchically and then weighted by their importance to the decision-maker. The total value of the alternatives is then calculated from the weighted sub-criteria scores. The procedure followed gives a two level hierarchy as illustrated in Figure 2.

Scores for each survey are formed by evaluating each sub-dimension and multiplying by weights and then further multiplying by the higher level weights, these are then aggregated. This is depicted in the following equation:

$$Total\ Score = \sum_{i=1}^{8} W_i^{\,2} \sum_{j=1}^{n} W_j^{\,1} S_j^{\,1}$$

Where W^1, W^2 and S^1 represent the weights for the first level, the weights for second level and the scores for the first level respectively.

In order to use the score tree to obtain the overall score and the scores for the first level, weights for the first level and sub-weights for the second level have to be assigned by the level of importance in each of the dimensions. Based on the idea of MCDM and as a pilot study, a focus group of five experts in survey research from the staff at Edinburgh Napier University was formed to construct a hierarchy of criteria and using Analytical Hierarchy Process to determine initial weights and sub-weights for all the quality dimensions. After several round meetings of discussions, the experts reached an agreement and the two level hierarchical criteria for assessing the quality of different stages or processes of a survey and a survey as a whole were obtained.

Criterion by criterion, the group assigned the weights and sub-weights to them by the level of importance in each of the stages or processes and balancing different choices or courses of selection with MCDM method. Comments on the figures derived in the pilot study, the final and full decision of the scores and weights was obtained from an online survey conducted on the members (about 500) of the European Survey Research Association (ESRA).

The online survey was carried out inviting all members of the ESRA to participate. The purpose of this survey questionnaire is to design an audit tool software giving a holistic assessment of survey quality. The idea of the assessment model is to score elements of the survey process in 8 dimensions or processes (as defined in the previous section such as formulation, sample frame, instrument, administration, data entry/validity, quality assurance, report and dissemination) and multiply the scores by the weights (importance). For each of the key dimensions (including all the eight first level dimensions, the second level dimensions for sample frame and instrument), a series of questions were asked or a set of criteria was applied and then weights and sub-weights were assigned for each question or criterion. The criterions are listed in Table 1.

Calibrating the assessment model

Out of the roughly 500 members of the ESRA, 110 questionnaires were completed, giving a response rate of 22%. There were 33% of the respondents with less than 9, 55% with 10-20 and 12% with more than 20 years of experience of designing, conducting and/or analysing surveys. There were 21% in the junior, 37% in medium and 42% in senior positions in their organisations. There were 80% from Europe, 12% from USA and 8% from the rest of world.

Taking into account that people's opinion may vary depending on the different types of surveys they usually dealt with, the questionnaire was designed to obtain viewpoints on weights for each of the eight domains in official, commercial and academic surveys separately. Analysing the results from the survey using one-way analysis of variance (ANOVA) indicated that there were no significant differences between different types of surveys from people's opinions on weights. In addition, the results also showed that the experts' opinion on weights did not depend on the respondents' experience with survey fieldwork, position in their organisations and location where they were from. Therefore the mean value on weights for the three types of survey was applied in the scoring tree of assessing survey quality as shown in Table 2. Along with the means, the values of standard errors were also calculated for all the eight first level dimensions, the second level dimensions for sample frame and instrument. As can be seen from Table 2, the values of standard errors are generally very small indicating consensus perceptions on the weights achieved from consistency amount of the respondents.

The Delphi method was employed and a panel of experts in survey research was formed from members of the ESRA to determine the final weights and sub-weights for the key dimensions in this research. Since the values of standard errors are generally very small, one can assume agreement between the experts in the study in the Delphi method phase 1, no further waves of the method was required; hence the mean weights are used in the score tree branches (see Figure 1 and Table 2).

The online survey targeted a group of experienced people who usually conduct survey assessment procedure. Within the proposed assessment system they assigned a percentage weight for each criterion or dimension from a range of 0 to 100 percent under the condition of summing all the weights to 100 percent at each level according to their professional judgement. In this way, the weights for the first level and sub-weights for the lower levels were obtained by the level of importance in each of the dimensions. For example, from formulation, sample frame, instrument, administration, data entry, quality assurance, report to dissemination, the weights for the first level were assigned from 0.16, 0.15, 0.18, 0.10, 0.11, 0.11, and 0.11 to 0.09; the sub-weights for the second level criteria assessing formulation were given from 0.23, 0.25, 0.20 and 0.32 and so on along with the standard errors as displayed in Table 2 below (for detailed description see Chen and Raeside, 2008).

The model is now applied to a number of surveys which have been compiled into a database by the researchers. This is done with a view to ascertain if the model can discriminate between surveys based on the dimensions of quality.

3. Classification analysis of a survey database

Data Collection

A survey database was formed from 103 regional, national and international surveys which were conducted between 1981 and 2009. The Information was collected from the Economic and Social Data Service (ESDS) database and UK Data Achieve (now UK Data Service) online, research reports, books and PhD theses in libraries. In the database some attributes of the surveys, such as survey mode, instrument, administration and dissemination etc., were examined. Twenty seven variables were used to profile the surveys; these detailed information of the surveys were taken and input into the database shown in Table 2.

The surveys in the database were all scored by the expert panel at Edinburgh Napier University using the quality assessment model according to the information collected, auditing the relevant detailed documents and their professional judgement. This generated quality assessment scores for all the surveys in the database overall and also in eight processes (dimensions) from formulation, sample frame, instrument, administration, data entry/validity, quality assurance, report to dissemination.

4. Data Analyses

Database characteristics

Within the formed survey database, there were 23 (22%) of the surveys which were hosted or sponsored by universities, 45 (44%) by government, and 35 (34%) by other organisations (such as Economic and Social Research Council, The Law Society, Transport research laboratory and Gambling Commission etc.). Universities executed 53 (51%), government conducted 14 (14%), and other organisations carried out 36 (35%) of the surveys. By comparison, most of the surveys in the database were hosted or sponsored by government departments and executed by universities.

As shown in Table 3, out of the 103 surveys in the database, there were 31.1% of them from Scotland, 51.5% from the rest of the UK, 7.8% from Europe, and 9.7% from the rest of the world. Thus the vast majority, altogether 82.6% of the surveys in the database are within the area of the UK. In comparison, the other organisations conducted the biggest proportion of the surveys from Scotland (38.9%); government departments carried out most of the surveys for the rest of the UK (64.3%); and universities executed the majority of the surveys for Europe (11.3%) and the rest of the world (17.0%).

Considering the methods of data collection, 41.7% of the surveys used face-to-face interviews; 39.8% of them were distributed by post; these two modes together account for 81.5%, and are the two main methods utilised in the data collection process. It is followed by telephone interviews which accounts for 10.7%, and 6.8% of the surveys employed web/email collection and only one utilised the other (actually a diary) method (1%). Compared by the type of organisation executed the surveys, government tends to do more face-to-face surveys, the other organisations are more likely to carry out telephone surveys, universities preferred the postal and web/email surveys.

Regarding the survey administration, as is displayed in Table 3, when the surveys were executed by government, every instrument was pre-tested; when the surveys were conducted by other organisations, almost all of the instruments (97.2%) were pre-tested; in both circumstances these organisations did very well in respect of pre-testing the survey instruments. However, when the surveys were carried out by universities, about 81.1% of the instruments were pre-tested, which was poor in comparison with the others. Overall, 89.3% of the instruments were pre-tested.

When the surveys were executed by government, 78.6% of the surveys were followed up. This indicates that the government is more likely to have follow-up procedures in place, whereas universities are less likely to do so (only 37.7%); other organisations are in between (55.6%). These may be because government has a propensity for doing more frequent surveys, i.e. they run the same survey again and again every one year or two only with minor amendment; while universities tend to do more one off surveys, e.g. research projects and PhD programmes. Overall, only nearly a half (49.5%) of the surveys had follow-up procedures.

Governments carried out most non-response analyses on the surveys (71.4%); both other organisations and universities conducted this less (50.9% and 50.0% respectively). This gave an overall total of 53.4% doing non-response analysis for all the surveys in the database.

For the types of dependent measures (the forms from which the survey measures depend on), 88.3% of the surveys were based on self-reporting, 10.7% on self-assessment and only one on the other dependent measures (actually a recorded description of the behaviour and activities). Thus, self-reporting is the dominant form of the dependent measures for all the types of organisations who executed the surveys in the database. With no exception, government surveys are all based on self-reporting while some of the universities and other organisations depend on self-assessment.

Of the organisations who executed the surveys (see Table 3), universities are more likely to carry out smaller surveys at lower cost (60.4% of them cost under 0.1 million pounds sterling, and only 7.5% more than one million) and involve a smaller number of people (54.7% of them have less than five people, and also only 7.5% more than 100). On the other hand, the government tends to run bigger surveys with higher costs (64.3% of them are more than 1 million pounds sterling and involve a larger number of people (50.0% of them have more than 100 people, and 21.4% between 20 and 100). In comparison, the other organisations are in the middle between the universities and the government in terms of the size (measured by the level of cost and the number of people involved in conducting the surveys) of the surveys.

Most of the surveys are of a smaller than bigger size in the survey database. In respect to the level of cost, 42.7% of the surveys incurred cost of less than 0.1 million pounds sterling, 33.0% had a cost of 0.1 to one million and only 24.3% more than one million. Likewise, in terms of number of people involved in conducting the surveys, 39.8% of the surveys had less than 5 people, 22.3% had 5 to 20, 17.5% had 20 to 100 and only 20.4% had more than 100 people.

When government executed the surveys, the annual surveys are the highest proportion (57.1%); whereas one off surveys are most common in universities and other organisations. Government and other organisations tend to do more frequent surveys than universities. In total, the proportion for annual surveys is 12.6%, once every 2-5 years 12.6%, occasionally 6.8% and one off 68%; therefore the one off survey is the most dominant frequency overall.

For the dissemination methods, universities are more likely to use traditional methods e.g. paper (56.6%), while most of the government (92.9%) and other organisations (58.3%) tend to apply the combination of modern methods such as web and traditional methods such as paper. However, for all the surveys in the database, 44.7% of the surveys disseminated their results with paper only, and 55.3% used both paper and web together.

In terms of impact of the survey results, as expected, universities mainly had an effect on academia, whereas government and other organisations had more influence on regional policy and the general public. In total, 36.9%, 44.7% and 18.4% of the survey findings influenced academia, regional policy and the general public, and national policy and the general public respectively.

The surveys were profiled using the assessment model and these generated 15 continuous measurements, to which one-way ANOVA was applied to ascertain if there were significant differences in these organisational types approached surveys. The first eight of these measures related to the quality dimensions and all showed significant differences between the organisational types at the 5% level. The nature of the differences in the quality dimensions are displayed in Figure 3. Overall the scores of formulation, sample frame, instrument, administration, report and dissemination are similar (scoring around 3.8) and are relatively higher than those of data entry/validity and quality assurance (which scored around 2.7). Considering the types of organisations conducting the surveys, universities performed best in terms of formulation, quality assurance and data entry/validity, however, they need to improve their performance on instrument, administration, dissemination and sample frame. Government did best in terms of sample frame, and need improve its performance on formulation, quality assurance and data entry/validity; and other organisations achieved better scores for instrument, administration, dissemination and report, and need improve their performance on quality assurance and data entry/validity.

The other measures obtained relating to assessment of the performance and complexity of the survey include response rate, number of respondents, number of questions, mean question length (number of words), percentage of closed

questions, and ease of coding and the ease of completion (measured with a five point Likert scale). The scores are displayed in Table 4, one-way ANOVA tests showed that the differences are significant between response rate, number of respondents, ease of coding, ease of completion and the percentage of the closed questions between the three types of the organisations conducting the surveys. However, average length of question instruction words is not found to be significant.

On average, university surveys had a significantly lower response rates and lower percentages of the closed questions, a smaller number of respondents than those of government and other organisations. There is no significant difference in the scores between government and other organisations. Government surveys have a significantly larger number of questions than those of universities and other organisations (there is no significant difference in the surveys between universities and other organisations).

Of all the surveys in the database, government surveys are significantly more difficult than those of the universities and other organisations with respect to both coding and completion (there is no significant difference in the surveys between universities and other organisations). The average level of difficulty is two for all the surveys if marking it from one (easy) to five (difficult). Therefore, in total there are most surveys were easier to complete and code than those which are relatively tougher to do so.

It is clear from the bar chart of the eight dimensions of survey quality in Figure 3 that there is no substantial difference between government and other organisations because most of them are similar; nevertheless, universities do appear to differ from both of them. Universities outperformed government and other organisations with regard to formulation, data entry/validity, quality assurance and survey quality as a whole; whereas government and other organisations were better than universities in respect to administration, instrument, sample frame, report and dissemination. However, for report, it is a different story - the survey quality from other organisations is slightly higher than that from universities, which is higher than that of government.

When the surveys were conducted by different organisations, the survey quality is higher by average from university, but it is not significant (p-value = 0.309). When surveys were hosted or sponsored by different organisations, the quality of the surveys was significantly different (p-value of 0.032). For example, the quality from government hosted surveys appears to be significantly higher (an average difference of 0.12) than those from universities (p-value of 0.024); but there is no significant difference in the average quality scores between the surveys from universities and other organisations.

Comparing the different regions, the survey quality in Scotland is the highest, followed by the rest of the UK, Europe, and the rest of the world scores the lowest. A one-way ANOVA with post hoc tests shows that there is a significant difference between the survey quality in Scotland and that in the rest of the world (an average difference of 0.18, p-value = 0.043); There was no significant difference between other regions. When comparing survey quality for the different methods of administration, face-to-face scores highest, followed by postal, and that of the other methods including telephone, the web/email, and other scores lowest. However, a one-way ANOVA test shows that there is no significant difference between them (p-value of 0.851).

The results in Table 5 also show that follow-up procedures and carrying out nonresponse analysis did improve survey quality but not significantly, (p-values for independent samples t-tests are 0.677, 0.298 and 0.707 respectively). Considering the type of dependent measures, the surveys that depended on self-reporting led to better quality sores than those that depended on self-assessment and other measures. However, this difference was not statistically significant (p-value of 0.944). Smaller costing surveys tend to have a higher quality as the mean survey quality scores with the low cost (less than 0.1 million) and medium levels of cost (between £0.1m to £1m) are higher than those with high level of cost (more than £1m) but not significantly so (p-value = 0.952).

Regarding the number of people involved in conducting the surveys, it is similar story as the level of survey cost. The surveys have highest quality when the medium number of people conducting them is between 5 and 20, followed by a small group of people (less than 5), and the surveys involving larger groups of people (20 or over) have the comparatively lowest quality. However, a p-value for one-way ANOVA of 0.432 indicates that there is no significant difference. For the frequency of the surveys, the quality is higher with one-off surveys than the other surveys which are frequent such as annual, occasional and the surveys conducted once every 2-5 years. Again the difference is not significant (p-value = 0.112).

The surveys disseminated with paper are better than those using a combination of both web and paper. The results are significant for an independent samples t-test, (average difference of 0.11, p-value = 0.03). National government surveys are better in respect of quality than those conducted by academia, however, surveys for academic research are better than those bringing impact to a region. But the differences are not statistically significant (p-value = 0.108)

From Table 6 it is indicated that there is a significant negative correlation between the time of the survey conducted and survey quality. This suggests that the survey quality has deteriorated over time. Survey quality is positively correlated to response rate and percentage of closed questions; while it is negatively related to the number of people involved in conducting the surveys, number of respondents, number of questions in the questionnaires, the maximum number of questions someone could complete, the percentage of questions requiring calculation and average length of question instruction words. However, none of the correlations were significant.

The significant correlations between the other variables are as follows:

- Response rate is positively correlated with number of people involved in conducting surveys and negatively correlated with number of respondents and the number of questions in the questionnaires.
- The number of people involved in conducting surveys is positively correlated with number of respondents, number of questions in the questionnaires, the maximum number of questions someone could complete and the percentage of questions requiring calculation.
- Number of respondents is positively correlated with number of questions in the questionnaires, the maximum number of questions someone could complete and the percentage of questions requiring calculation.
- Number of questions is highly positively correlated with the maximum number of questions someone could complete and the percentage of questions requiring calculation.
- The maximum number of questions someone could complete is highly positively correlated with the percentage of questions requiring calculation.
- Percentage of closed questions is highly positively correlated with average length of question instruction words.

There are no other significant correlations between the variables. Though the above correlations are significant, the only strong correlation (the Pearson's correlation coefficient = 0.964) is between the variables of the maximum number of questions someone could complete and the percentage of questions requiring calculation. All the other correlations are not strong (the Pearson's correlation coefficients are all below 0.5). In addition, there may be some problems with the Pearson's correlation analysis, because, some of the relationships between the variables may not be linear.

5. Conclusion

Research has been undertaken to improve the quality and effectiveness of surveys and from reviewing this work a model to assess the quality of surveys has been developed. The model has been applied to over 103 regional, national and international surveys in this paper. This generates data on formulation, design and application quality. These data have been used along with attributes of the surveys examined to create a database. Various measures of a survey were taken such as average question length, number of questions and entered into the database. This database has been analysed to identify the structural features of surveys that are associated with survey quality.

When the quality scores for eight survey processes are compared, there are two groups that can be identified; those of formulation, sample frame, instrument, administration, report and dissemination (they are similar to each other and together form one group) are relatively higher than those of data entry/validity and quality assurance (both of them are alike and form another group). In respect of the types of organisations conducting the surveys, universities outperformed government and other organisations in terms of formulation, quality assurance and data entry/validity, government and other organisations performed better in terms of administration, instrument, sample frame and dissemination. However, universities outperformed overall to government and the other organisations in respect of survey quality as a whole, though the quality of the surveys carried out by universities are more variable. There appears to be no significant difference in the quality of the surveys conducted by other organisations and government.

Other findings from the analysis are as follows:

- The quality of surveys from government hosted is significantly higher on average than those hosted from universities;
- The surveys disseminated with paper are significantly better on average than those using a combination of both web and paper in our survey database; and
- The quality of surveys deteriorates significantly over time.

From this, it is recommended that: Universities, when conducting surveys, should make effort to increase response rates, gain funding for research, concentrate on the sampling frame, instrument, administration and dissemination processes of surveys. Government and the other organisations should employ universities to execute surveys, reduce the number of

questions included in surveys, and pay more attention to the processes such as data entry/validity and quality assurance of surveys. More and more surveys are carried out and more and more information are collected globally, but the quality of the data is uncertain, therefore not only every effort should be made to improve survey quality but on every occasion users should be aware of that the data used might not be trustworthy. There should be continuous holistic assessment of the validity and reliability of the data before it is used.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which helped to improve the quality of the paper.

References

Alwin, D. (2007). *Margins of error: a study of reliability in survey measurement*, Wiley-Blackwell.

Andersen, R., Kasper, J. and Frankel, M. (1979). *Total survey error*, Jossey-Bass.

Andersson, C., Lindstrom, H. L. and Lyberg, L. (1997). *Quality declarations at Statistics Sweden—principles and practices*. Washington DC., Statistical Policy Working Paper 26. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

Arondel, P. and Depoutot, R. (1998). *Overview of Quality Issues when Dealing with Socio-Economic Products in the International Environment*. Paper presened at the XXXth ASU Meeting.

Assael, H. and Keon, J. (1982). *Nonsampling vs. sampling errors in survey research*. *The Journal of Marketing* 46(2): 114-123.

Belton, V. and Stewart, T. (2002). *Multiple criteria decision analysis: an integrated approach*, Springer.

Bergdahl, M., Black, O., Bowater, R. and Chambers, R. (1999). *Model Quality Report in Business Statistics*." *vol. 1, Theory and Methods for Quality Evaluation. Luxembourg: Eurostat.*

Bound, J., Brown, C. and Mathiowetz, N. (2001). *Measurement error in survey data*. *Handbook of econometrics* 5: 3705-3843.

Burg, T. (2004). *Quality Reports at Statistics Austria*. European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz, Germany.

Carson, C. (2011). *Toward a framework for assessing data quality*, International Monetary Fund.

Chen, T. (2011). *Measurement and Assessment of Survey Quality: The Development of a Holistic and Quantitative Tool*. ISBN: 978-3-8443-8461-1. Lambert Academic Publishing.

Chen, T. and Raeside, R. (2008). *Assessing the Quality of a Survey*. *Edinburgh Napier University, School of Accounting, Economics & Statistics. | 2008 | ISBN: 9781873869772.*

Dalenius, T. (1977). *Bibliography on Non-Sampling Errors in Surveys: I (A to G)*. International Statistical Review/Revue Internationale de Statistique, Longman Group Ltd. 45: 71-89.

Dalkey, N. C. (1969). *The Delphi method: An experimental study of group opinion*, Rand Corp. Santa Monica, CA.

Eurostat (2002). *Quality in the European statistical system - the way forward*. *Eurostat, Luxembourg. (Available from http://europa.eu.int/comm/eurostat)*.

Eurostat (2003). *Definition of Quality in Statistics*. *Eurostat, Luxembourg. (Available from http://Eurostat/A4/Quality/03/General/Definition).*

Groves, R. (2004). *Survey errors and survey costs*, Wiley-Interscience

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E. and Tourangeau, R. (2009). *Survey methodology*, John Wiley & Sons Inc.

Laiho, J., Hietaniemi, L. and Jonasson, T. (2002). *Quality Guidelines for Official Statistics*, Statistics Finland.

Marriott, F. and Kendall, M. (1990). *A dictionary of statistical terms*, Longman Scientific & Technical.

Moser, C. and Kalton, G. (1979). *Survey Methods in Social Investigation,* Heinemann Educational Books, London.

OECD (2003). *Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities*. *Available from http://www.oecd.org/dataoecd/26/38/21687665.pdf*.

O'Muircheartaigh, C. (1997). *Measurement error in surveys: A historical perspective*. Survey measurement and process quality: 1-25.

ONS (2004). *Guidelines For Measuring Statistical Quality. London: Office for National Statistics.*

Osborne, J. (1942). *Sampling errors of systematic and random surveys of cover-type areas.* Journal of the American Statistical Association 37(218): 256-264.

Saaty, T. (1985). *Decision Making For Leaders: Life Time Learning Publications. Belmont, California*.

Schwarz, N., Groves, R. M. and Schuman, H. (1995). *Survey methods*. *SMP Working Paper Series, no.30. Preliminary Chapter for Handbook of Social Psychology. D. Gilbert, et al. (Eds.). New York: McGraw Hill.*

Statistical Policy Office USA (2001). *Measuring and Reporting Sources of Error in Surveys*. *STATISTICAL POLICY WORKING PAPER 31,Office of Information and Regulatory Affairs, Office of Management and Budget, Washington DC USA*.

Statistics Canada (1998). *Statistics Canada Quality Guidelines*. *3rd edition, Catalogue no. 12-539-XIE, p. 101. Ottawa: Statistics Canada.*

Tolonen, H., Helakorpi, S., Talala, K., Helasoja, V., Martelin, T. and Prättälä, R. (2006). *25-year trends and socio-demographic differences in response rates: Finnish adult health behaviour survey*. European journal of epidemiology 21(6): 409-415.

Weisberg, H. (2005). *The total survey error approach: A guide to the new science of survey research*, University of Chicago Press.

Table 1: Dimensions used to assess surveys and weights assigned

Dimension and subcategory	Weight	Standard error
1.Formulation	0.16	0.0072
Community involvement	0.23	0.0079
Academic involvement	0.25	0.0077
Policy Importance	0.20	0.0065
Development	0.32	0.0087
2.Sample Frame	0.15	0.0060
Failure to contact some subjects	0.31	0.0139
Coverage	0.47	0.0110
Incomplete responses	0.22	0.0101
3.Instrument	0.17	0.0089
Questionnaire	0.4	-
Specification problems	0.20	0.0090
Question wording	0.21	0.0094
Length of the questions	0.11	0.0052
Length of the questionnaire	0.14	0.0042
Order of questions	0.11	0.0043
Response categories	0.12	0.0045
Questionnaire format	0.11	0.0049
Respondent	0.4	-
Comprehension	0.37	0.0138
Retrieval from memory	0.20	0.0067
Judgment	0.19	0.0078
Communication of response	0.24	0.0085
Harmonization	0.2	0.0031
4.Administration	0.10	0.0044
5.Data Entry/validity	0.11	0.0042
6.Quality Assurance	0.11	0.0048
7.Report	0.11	0.0059
8.Dissemination	0.09	0.0057

Figure 1: The survey process flow chart (Chen, 2011)

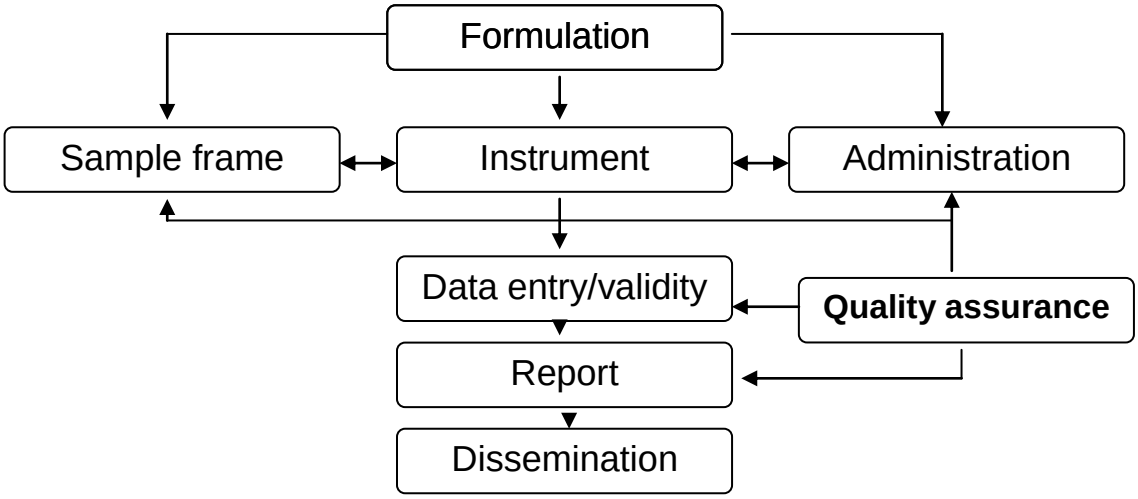

Figure 2: The scoring tree of assessing survey quality (Chen, 2011)

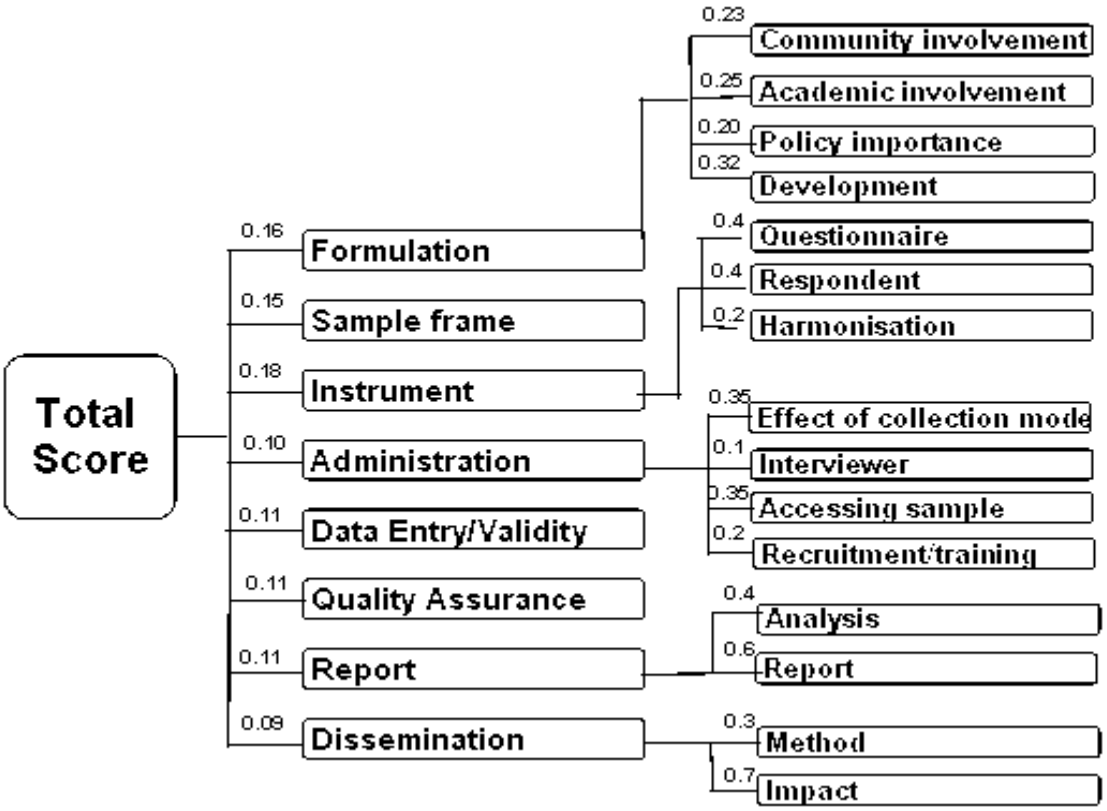

Figure 3: The quality of eight survey processes by survey organisations

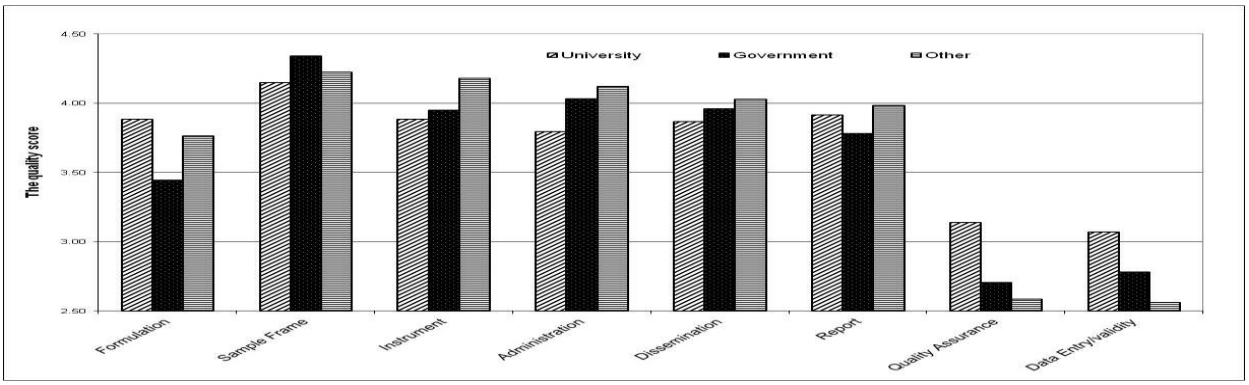

Table 2: The information collected in the survey database

1. The name of the survey
2. The date of fieldwork or time period covered by the survey
3. The sponsor(s)/organisation(s) who commissioned the survey
4. The principal investigator(s)/organisation(s) who conducted the survey
5. The purpose or aim of the survey
6. The country, region or location of units of observation for the survey
7. The method of data collection
8. The survey population
9. The survey response rate
10. Whether the survey instrument was pre-tested or not before the actual fieldwork
11. The follow-up procedures after the actual fieldwork
12. Whether the a non-response analysis for the survey was carried out or not
13. The type of dependent measures for the survey
14. The survey cost
15. The number of people involved in the survey fieldwork i.e. the management and organisation team and interviewers etc.
16. The number of people who responded to the survey
17. The time dimensions of the survey
18. The number of questions in the survey
19. The maximum number of questions someone could complete
20. The percentage of closed questions
21. The percentage of questions requiring calculation
22. Average length of question instruction words
23. Ease of completion, marked1 to 5 from easy to difficult
24. Ease of coding, marked 1 to 5 from easy to difficult
25. The frequency of the survey
26. The dissemination method(s) of the survey results
27. The impact(s) of the survey findings

Table 3: The categorical variables by survey organisations

Variables	Type of organisation executed survey %			
	University (N=53)	Government (14)	Other (36)	Total (103)
Region:				
Scotland	26.4	28.6	38.9	31.1
Rest of UK	45.3	64.3	55.6	51.5
Europe	11.3	7.1	2.8	7.8
Rest of World	17.0	0.0	2.8	9.7
Method:				
Face-to-face	32.1	64.3	47.2	41.7
Telephone	5.7	7.1	19.4	10.7
Postal	50.9	28.6	27.8	39.8
Web/email	9.4	0.0	5.6	6.8
Other	1.9	0.0	0.0	1.0
Pre-test of instrument:	81.1	100.0	97.2	89.3
Follow-up procedures:	37.7	78.6	55.6	49.5
Nonresponse analysis:	50.9	71.4	50.0	53.4
Type of dependent measures:				
Self report	86.8	100.0	86.1	88.3
Self assess	11.3	0.0	13.9	10.7
Other	1.9	0.0	0.0	1.0
The level of cost:				
Low<0.1m	60.4	0.0	33.3	42.7
Medium 0.1-1m	32.1	35.7	33.3	33.0
High>1m	7.5	64.3	33.3	24.3
Number of people involved:				
Small<5	54.7	0.0	33.3	39.8
Medium 5-20	26.4	28.6	13.9	22.3
Large 20-100	11.3	21.4	25.0	17.5
Very large >100	7.5	50.0	27.8	20.4
Frequency:				
Annual	1.9	57.1	11.1	12.6
Once every 2-5 years	5.7	14.3	22.2	12.6
Occasionally	3.8	21.4	5.6	6.8
One off	88.7	7.1	61.1	68.0
Dissemination methods:				
Web & paper	43.4	92.9	58.3	55.3
Paper	56.6	7.1	41.7	44.7
Impact:				
Academia	49.1	0.0	33.3	36.9
Regional policy & public	30.2	92.9	47.2	44.7
National policy & public	20.8	7.1	19.4	18.4

Table 4: The continuous variables by survey organisations

	University			Government			Other			P-value of ANOVA
	N	Mean	Std. Error	N	Mean	Std. Error	N	Mean	Std. Error	
Response Rate (%)	53	42	4	14	65	4	36	57	4	0.001
Number of respondents	53	2627	830	14	30671	13460	36	10023	2365	0.001
Average length of question instruction words	53	29	9	14	24	2	36	21	1	0.686
Ease of coding mark 1 (easy) to 5 (difficult)	53	2	0	14	3	0	36	2	0	0.002
Ease of completion mark 1 (easy) to 5 (difficult)	53	2	0	14	4	0	36	2	0	0.001
% closed questions	53	78	4	14	94	1	36	93	2	0.006
Number of questions	53	81	38	14	357	75	36	80	14	0.001

Table 5: Overall Survey quality

Categorical variable with p-value (in brackets) from statistical test	Survey quality as a whole					
	N	Mean	Std. Error	Median	Min	Max
Type of organisation hosted (0.032):						
University	23	3.62	.06	3.68	2.95	3.99
Government	45	3.74	.02	3.74	3.49	3.97
Other	35	3.72	.03	3.72	3.22	4.12
Type of organisation executed (0.309):						
University	53	3.76	.03	3.74	2.95	4.12
Government	14	3.67	.03	3.68	3.48	3.93
Other	36	3.69	.03	3.70	3.22	4.00
Region (0.047):						
Scotland	32	3.77	.03	3.76	3.46	3.99
Rest of UK	49	3.70	.02	3.71	3.21	4.00
Europe	12	3.66	.08	3.65	3.40	3.96
Rest of World	10	3.59	.11	3.67	2.95	4.12
Method (0.851):						
Face-to-face	43	3.73	.02	3.73	3.51	4.00
Postal	41	3.71	.04	3.72	3.13	4.12
Phone, web and other	19	3.69	.06	3.72	2.95	3.96
Pre-test of instrument (0.677):						
Yes	92	3.72	.02	3.73	2.95	4.12
No	11	3.68	.02	3.69	3.60	3.77
Follow-up procedures (0.298):						
Yes	51	3.73	.02	3.75	3.21	4.00
No	52	3.69	.03	3.72	2.95	4.12
Non-response analysis (0.707):						
Yes	55	3.71	.02	3.79	3.40	4.12
No	48	3.70	.04	3.71	2.95	4.00
Type of dependent measures (0.944):						
Self report	91	3.71	.02	3.73	2.95	4.12
Self assess and other	12	3.69	.07	3.73	3.21	3.96
The level of cost (0.952):						
Low<0.1m	44	3.71	.04	3.75	2.95	4.12
Medium 0.1-1m	34	3.71	.03	3.73	3.30	3.97

High>1m	25	3.70	.02	3.71	3.48	3.93
Number of people involved (0.432):						
Small<5	41	3.70	.04	3.75	2.95	4.12
Medium 5-20	23	3.76	.03	3.81	3.49	3.97
Large 20-100	18	3.68	.03	3.70	3.30	3.86
Very large >100	21	3.67	.02	3.69	3.48	3.85
Frequency (0.112):						
One off	70	3.74	.03	3.77	2.95	4.12
Other	33	3.64	.02	3.65	3.30	3.86
Dissemination method (0.03):						
Web & paper	57	3.65	.02	3.69	3.22	3.87
Paper	46	3.76	.04	3.85	2.95	4.12
Impact (0.108):						
Academia	38	3.73	.04	3.84	2.95	3.99
Regional policy & public	46	3.66	.02	3.69	3.22	3.87
National policy & Public	19	3.77	.03	3.72	3.57	4.12

Table 6: Correlation matrix for the continuous variables

Variables	Response Rate (%)	Number of people involved	Number of respondents	Number of questions	% closed questions	% questions requiring calculation	Average length of question instruction words	Survey quality as a whole
Response Rate (%)	-.051	.052	.019	.112	.110	.105	-.039	-.209*
Number of people involved		.399**	-.263**	-.198*	.008	.111	-.070	.023
Number of respondents			.439**	.455**	.111	.238*	.005	-.075
Number of questions				.234*	.119	.209*	-.022	-.103
% closed questions					.126	.300**	-.027	-.135
% questions requiring calculation						.109	.352**	.069
Average length of question instruction words							-.011	-.010
Survey quality as a whole								-.025

*Note: * significant at 0.05, ** significant at 0.01.*