

Survey as a source of low quality research data

Grzegorz Szyjewski
University of Szczecin, Faculty of
Economics and Management,
ul. Mickiewicza 64, 71-101
Szczecin, Poland
Email:
grzegorz.szyjewski@usz.edu.pl

Luiza Fabisiak
West Pomeranian University of
Technology, Szczecin.
ul. Żołnierska 49, 71-210
Szczecin, Poland
Email: l.fabisiak@wi.zut.edu.pl

Abstract—Survey is the most common way to gather information for research purposes. Gathering information is usually provided in the early stage of research procedure. The information create the basis for further activities which lead to scientific research results. The problem is that information gathered as a result of survey is exposed to the high risk of errors. Following factors generates mostly errors: the way of survey carrying out, the survey content and just by simple human factors. Researchers very rarely pay their attention to the data quality taking as an obvious that gathered information describes properly studied case. In the end of the process they get results, which are becoming basis for the conclusions. Taking that into account the questions is about the quality, of research results, which were based on the surveys. What is important, usually most publications do not contain any information about source data on the basis of which results came out. It disables the independent evaluation of the results. The goal was to investigate the scale of survey data errors, using the scientific experiment. To achieve it, authors used their own survey system, which helped them in instant verification of the respondents' answers.

I. INTRODUCTION

SURVEY is probably the most popular method of data gathering for research purposes [1]. It is used as a supporting tool not only for science but for political and business aspects of life as well [2] [3] [4]. Surveying consists in gathering data from respondents using dedicated tool. Data collected that way should form the standardized set, mapping reality being examined. People to be surveyed form the target of previously prepared questions. Questions can refer to themselves or to theirs' opinions about the subject being examined [5]. During survey the answers are being collected into a set, that forms the source data, which is a basis for the further steps of research procedure. Using appropriate research methods [6] previously gathered data is then analyzed.

The issue that can occur is big possibility of errors in gathered set of data. The most important is that even single errors in gathered data can have negative effect on all set [7]. The low quality of the set, of the source data, affects the further research procedure. That can cause wrong analysis results and false complex process results. It is important to mention that a mistake made in the beginning phase of

research procedure, becomes more difficult to be discovered while executing the furthers steps. The next problem is lack of possibility to undermine answers given during survey. That is because the answers are individual convictions of the respondents, who represent tested reality [8] [9]. It is difficult to prove that answers given by respondents are wrong and that they do not represent theirs real preferences.

Considering the above-mentioned issues, there comes the question about the quality of source data that is used as a basis for the scientific research. As we can find in the different sources, the problem of errors in source data is known and is being current since many years. Assuming that most research is conducted on the basis of such data sets, their results may be incorrect.

II. THE ANALYSIS OF THE PRESENT SITUATION

The problem of the errors in surveys' data and low quality of data are being mentioned since long time. Unfortunately large group of researchers don't pay appropriate attention to the quality of used by them data. This is also confirmed by studies published in 1996 [10]. The results of these studies suggest that errors in the source data are not perceived as a significant factor affecting their poor quality. Scientists often ignore the issue of source data errors.

There are many proven causes of low quality survey data. Such a situation is caused by many factors [11] [12], which are often ignored by the survey authors. Unfortunately, most of them cannot be eliminated if the survey data collection method is used. Rapid development of ICT, caused that standard surveys are replaced by electronic ones. Such technology made possible reaching faster more respondents. The main advantage of collecting data through electronic surveys is that data are ready for analysis, immediately after survey is finished [13]. The biggest disadvantage of online public surveys is the lack of a representative sample [14]. The interviewer can't verify the affiliation of his respondents to the desired target group. It seems that optimal way is to conduct electronic surveys, under the supervision of an interviewer. The group of respondents can be verified and collected data are stored into the database.

In data quality research, attention is also paid to the lower involvement of the respondent, in the case of a web survey [15]. The problem with the respondent's involvement is also due to the volume of survey. Keeping the respondent's attention becomes more difficult, the longer the time it takes to respond. An interesting solution for that problem may be mobile technology where survey questions are sent periodically to the respondent's mobile devices [16]. That shows that the development of information and communication technologies significantly increases elimination of survey errors.

It is also very important, in the discussed subject, that the majority of published research results do not contain any source data. Verification of the result is almost impossible, without access to the source data that was used. Sharing data is a natural in custom areas such as: astronomy and genomics [17]. Publishing the source data together with the research results, gives the proof of data good quality [18]. Authors of those publications choose their research to be publicly verified. Publishing source data also gives opportunity for other researchers, to use them in their own work. It has also been proved to have a positive effect on the author's citation rate [19]. In science we can observe a trend in which sharing source data is promoted. Some researchers support that path and some don't [20] [21] [22] [23] [24][25].

III. THE EXPERIMENT

Present knowledge suggests that the data collected from the surveys very often contains errors. Using such a data cause that received results may be incorrect. The aim of carried out experiment was to diagnose the scale of the problem. The goal was to verify possible amount of errors in the set of data gathered from survey. During the acquisition of data for the research, an electronic survey was used. It was filled under the supervision of an interviewer. The specificity of the experiment required establishing own individual survey system. That is why own author's ICT system was prepared and used. This was due to necessity to enter data as a respondent's declaration and their immediate practical verification. This activity allowed to check the quality of declared responses. Immediate verification of the respondent's declarations was possible thanks to the QR code technology (Quick Response). QR code works similar to the bar code, which is the basis for automatic identification [26]. The graphic form of the code stores data that can be read by optical input devices, such as a digital camera.

The scenario of the survey assumed the following stages: the exclusion of respondents who don't have the technical capabilities to join the survey, placing the survey question and receiving the respondent's answer in a declaration form, the practical verification of the previous declaration. 216 persons participated in the experiment. People aged 19-23 were invited to the study. It was assumed that this age group consists the most people that should be able to solve the

problem presented in the study. The sample group included people connected with technical and non-technical studies fields.

After people who didn't have technical capabilities for survey participation were eliminated, 197 qualified for the second stage. That group has been tested and questioned with two questions concerning the same issue. The feedback coming from the first answer was in a form of declaration. That question contained the graphic shown in Figure 1 and the question: "You walk into the shop to buy some electronics and you see the poster shown above. Would you be able to get this discount?" The respondents declared whether they would be able to use the promotion. The second question was concerning the same issue, but feedback was based on the real use (scan) of the QR code. Correct scanning of the displayed code formed a practical verification, of the previously declared ability. The design of the author's survey system allowed to count how many people confirmed their declaration during the practice part of the test. Each respondent received individual code and only the real scan, saved the response to the survey database.

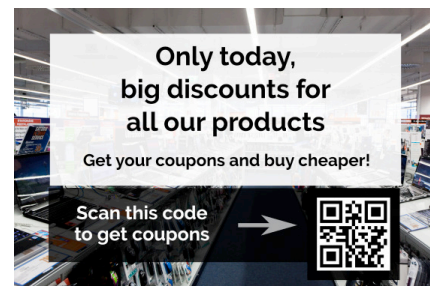


Fig. 1 Graphic of the survey question

The next step in the research procedure was to compare the answers given to both questions. That was a comparison of the declaration, with the real replies given by respondents. By comparing the results from both elements, you can discover possible scale of errors in respondents' declarations.

IV. DETERMINANTS OF FORMAL EVALUATION OF THE MULTICRITERIAL DECISION PROBLEM

An important phase was studying the error formation, based on data coming from examination of the preferences of a particular group. In order to eliminate factor of the "Digital exclusion", technical and economic fields students attended the study. In the research author used the method of multidimensional functional analysis - the scoring method [27]. The scoring method is based on a specific hierarchy of elements and their distance from the maximum possible achievable value. In the method, hierarchy is characterized by an increasing or decreasing, indicating the level of global criteria realization, which contains all sub-criteria [28]. Using the point method collects the information about criteria. The information receives assigned value according

to a comprehensive scale of values (Table 1) and the results are listed in the summary table [29].

It should be pointed that the criteria used in the elaboration are treated equal and values of factors are designated to the scale of preferences. In the assumed and accepted method, in assessing the preferences of the surveyed students, the following actions were considered:

- appointing an expert group of representatives of users and potential users of QR technology, selected from a rated group (university students);
- defining detailed criteria for QR codes assessment, their hierarchy and the relationships between the evaluation factors which means creating a detailed survey;
- compiling results in the form of individual tables and in the form of aggregate tables;
- final results analysis and conclusions (preferences of survey respondents and AHP decision support method)

In the point method, experts assign ratings to the criteria according to the scale consistent with the Table 1.

TABLE I.
SCORING RULE EVALUATION OF EACH OF THE TEST CRITERIA

Grades range	Explanation
0	Represents lack of a given feature
0,25	A low (satisfactory) level of a feature
0,5	An average (sufficient) level of a feature
0,75	High (fair) level of a feature
1	Exceptionally high level of a given feature

The point method that was selected for the assessment allows partial evaluation for each of tested criterion. That method does not specify the criterion as the worst / the best, it is only derived from the normalized distance. On the basis of the results coming from the point rating of respondents' preferences, the Analytical Hierarchy Process (AHP) method was used [30]. The relevance in the AHP method for transforming the scores obtained from the expert marks each criterion and factor in the hierarchical model, gained from the pairwise comparison. The result of comparing two elements from the same hierarchy level is reflected by the existing domination between them. For the domination description we use nine-step preference scale. Then all elements values in the row are summed [31] [32] [33]. The sums are normalized according to the formula (1) in further steps of the method we determine the coherence of the ratings (4), which is the same as the transitions of the criteria weights. It should be emphasized that in order to assess the validity of the criteria to be considered consistent, the value of the calculated conformity factor should not be greater than 0.1.

$$an_i = a_i / \sum_{n=1}^i a_n \tag{1}$$

The divergence factor is derived from the formula (2):

$$CI = \frac{|n - \lambda_{sr}|}{n-1} \tag{2}$$

Where: n is the number of criteria where (matrix rows) and λ_{sr} is the consistency factor.

The compliance indicator is derived from formula (3):

$$CR = \frac{CI}{R} \tag{3}$$

Where, CR is the compliance rate, CI is the coefficient of discrepancy and R is the factor random compliance.

The cohesion factor is determined using formula (4):

$$\lambda_{sr} = \frac{1}{n} \sum_{i=1}^n \lambda_i \tag{4}$$

Where, in λ_i is described by the formula (5):

$$\lambda_i = \frac{\sum_{j=1}^n A_{ij} * w_j}{w_i} \tag{5}$$

Where, w_i is the weight of the criterion i , A_{ij} is the matrix element A .

The coefficient of randomness of R is dependent on the number of criteria being taken into account. The value of this factor according to the number of criteria is shown in Table 2.

TABLE III.
VALUES OF CONFORMITY R FOR SPECIFIC QUALITY OF CRITERIA

n	1	2	3	4	5	6	7	8
R	0	0	0,52	0,89	1,11	1,25	1,35	1,4
n	9	10	11	12	13	14	15	
R	1,45	1,49	1,51	1,54	1,56	1,57	1,58	

Similarly, the consistency of the preferences of the choice options can be determined in relation to each criterion. We should adopt n not as the number of criteria but as the number of variants. The synthesis of the criteria importance and alternatives preferences according to each criterion consists in multiplying the weight of a given criterion by the value of the variant evaluation for that criterion. That action is done for all criteria, and the results received for each criterion are summed up. As a result, we obtain the generalized alternative quality measure. Those actions are repeated for each variant and then variant ranking is created. The highest quality value means the best considered variant [34]. As the result of the AHP method we receive the ranking of variants created according to the quality measure coming from each decision alternative. That method determines whether the user is able to use the QR code correctly.

V. EMPIRICAL VERIFICATION

The presented research is based on results of an online survey - more in Chapter III: The Experiment. Through the survey all phenomena related to the survey errors formatting,

based on data from the respondents' preferences were categorized. Results are classified by 4 characteristics (conditional criteria). The studies used variables, which had adopted integers value from the range [0-1]. Data characterized the status of variables as:

- K1 Are you using a mobile phone?
- K2 Do you use an Internet access in your mobile?
- K3 Entering the electronics store with a decision to make a purchase, you see the poster as shown below. Do you know what to do to get a discount?
- K4 Now, scan the code shown below and see what it is hiding.

It should be noticed that the choice of the criteria did not solve the problem of decision selection. It was necessary to determine the preference vector for those criteria. Due to the gained results, questions arise. What were the preferences of respondents who responded to the survey questions? What did they pay the most attention to? And did the students not cheat while answering survey questions? The last key question is whether respondents answered all questions correctly? Receiving an answer and determining the preference vector for the selected criteria requires a point method (single or group context). Results of the point method of scoring the preferences of respondents are presented in Table 3.

TABLE III.
ESTIMATION OF RESPONDENTS' PREFERENCES ACCORDING TO CRITERIA

	Preferences	Point scale	Explanation
K1	0,91	1	Very good level of features
K2	0,95	1	Very good level of features
K3	0,74	0,75	Good level of features
K4	0,25	0,25	Satisfactory level of features

After receiving ratings from the survey and pointing the respondents' preferences, the next step was AHP method use. The AHP method identifies in dialog mode, not very clear statements of a selected group of experts (respondents answering the survey questions). The characteristics of QR codes according to the evaluation criteria, indicated by the AHP method and the preferences of the respondents are shown in Table 4.

Table III.

QR CODES ACCORDING TO THE INDICATED EVALUATION CRITERIA

Criteria	K1	K2	K3	K4	Scales
K1	0,17	0,17	0,16	0,19	0,171
K2	0,18	0,18	0,16	0,19	0,179
K3	0,31	0,31	0,28	0,25	0,288
K4	0,34	0,34	0,40	0,37	0,362

In order to assign each criterion's weight according to the AHP procedure, the criteria were compared in pairs and the preference of the decision maker was assessed. The next step was the matrix of the relative criteria importance construction. Based on the matrix's own value, the generalized relevance of criteria were determine. Shown in Table 5.

TABLE V.
GENERALIZED VALIDITY OF CRYTERIA

Criteria	Sum of grades	Scales	λ_{max}
K1	5,82	0,171	1,00
K2	5,57	0,179	1,00
K3	3,57	0,288	1,03
K4	2,73	0,362	0,99

The AHP method uses the 1-9 scale. The goal is to find out the advantage of one variant over another, according to a particular criterion. Quantitative data were used in that study.

As a result of the AHP procedure, we received the W1 preferences matrix (declared data) and W2 (practically verified data), according to all criteria and containing generalized preferences. That matrix is shown in Table 6 and the distribution of results in Figure 2.

TABLE VI.
VARIANT PREFERENCE MATRIX

	K1	K2	K3	K4	Scales
W1	0,476	0,487	0,425	0,200	0,364
W2	0,524	0,513	0,575	0,800	0,636

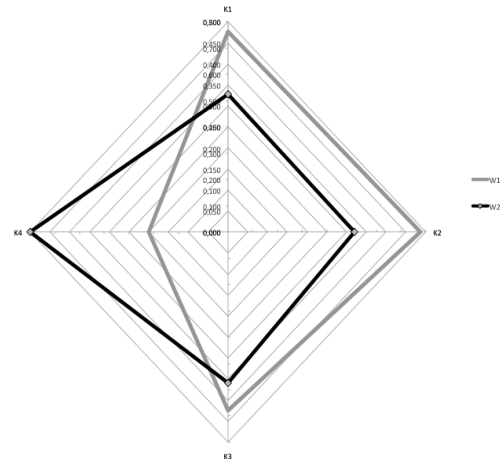


Fig 2 Variant Preference Matrix according variants W1 and W2

In the preferences matrix distribution according variants W1 and W2 shown above, we can observe clearly dominance of W2. The difference between variants 1 and 2 was the first element recorded standard survey declaration and the second one examined the actual situation. Thanks to the presented results received using author's survey system, we could investigate the phenomenon of survey data error formation. In the experiment we combined declared values with data practically verified. That study showed the probable scale of the low quality data coming from survey research. The collected data indicate that part of respondents were wrong about their competence in the examined subject. Thanks to the QR and other IT technologies that phenomenon was discovered and described in the above research.

VI. CONCLUSION

The result forms the basis for the analysis of the current respondents behaviour. It concerns the area of justification and the subsequent efficiency using survey as a method of obtaining data for research. The presented approach to multicriteria evaluation of the decision problem allowed us to use the point method for the evaluation of the respondents' preference designation and the AHP. Thanks to the used survey tool, the data obtained for the experiment include not only the declared responses, but also their practical verification. A comparison of the data from two sources showed that the theoretical data obtain to be significantly different from the real. Using such data sets in research may lead to wrong results and cannot be verified.

The current state of ICT development allows step by step resignation from survey data acquisition methods. We have presently more and more opportunities to use modern mobile devices and data-communication technologies for data collection [35]. Those technologies can be used successfully to conduct research procedures and they can eliminate respondents' involvement in the survey process.

REFERENCES

- [1] S. L. Pfleeger, B. A. Kitchenham, "Principles of Survey Research Part 1: Turning Lemons into Lemonade" in ACM SIGSOFT Software Engineering Notes, vol. 26, no. 6, 2001, pp. 16-18.
- [2] R. Libby, P. C. Fishburn, "Behavioral models of risk taking in business decisions: A survey and evaluation", in *Journal of Accounting Research*, 1977, pp. 272-292.
- [3] G. Enderle, "A worldwide survey of business ethics in the 1990s", in *Journal of Business Ethics*, vol. 16, no. 14, 1997, pp. 1475-1483.
- [4] E. Quintelier, S. Vissers, "The effect of Internet use on political participation an analysis of survey results for 16-year-olds in Belgium", in *Social Science Computer Review*, vol. 26, no. 4, 2008, pp.411-427.
- [5] A. Pinsonneault, K. Kraemer, "Survey research methodology in management information systems: an assessment.", in *Journal of management information systems*, vol. 10, no. 2, 1993, pp. 75-105.
- [6] R.K. YIN, *Case Study Research. Design and Methods*, Sage Publications, Thousand Oaks - London - New Delhi 2003, s. 5-9.
- [7] M. H. Hansen, W. N. Hurwitz, W. G. Madow, "Sample survey methods and theory" Vol. 1, p. 638. New York: Wiley, 1953
- [8] R. M. Groves, "Research on survey data quality." in *The Public Opinion Quarterly*, vol. 51, 1987, pp. S156-S172.
- [9] P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, S. Sudman, "Measurement errors in surveys" Vol. 173, John Wiley & Sons, 2011.
- [10] R. Y. Wang, D. M. Strong, "Beyond accuracy: What data quality means to data consumers", in *Journal of management information systems*, vol. 12, no. 4, 1996, pp. 5-33.
- [11] J. Bound, C. Brown, N. Mathiowetz, "Measurement error in survey data." in *Handbook of econometrics*, vol. 5, 2001, pp. 3705-3843.
- [12] F. M. Andrews, A. R. Herzog, "The quality of survey data as related to age of respondent", in *Journal of the American Statistical Association*, vol. 81, no. 394, 1986, pp. 403-410.
- [13] B. Hanscom, J. D. Lurie, K. Homa, J. N. Weinstein, "Computerized questionnaires and the quality of survey data.", in *Spine*, vol. 27, no.16, 2002, pp. 1797-1801.
- [14] C. Marta-Pedroso, H. Freitas, T. Domingos, "Testing for the survey mode effect on contingent valuation data quality: A case study of web based versus in-person interviews", in *Ecological economics*, vol. 62, no.3, 2007, pp. 388-398.
- [15] A. W. Meade, S. B. Craig, "Identifying careless responses in survey data.", in *Psychological methods*, no. 17, vol. 3, 2012, p. 437.
- [16] M. P. Couper, "Technology trends in survey data collection.", in *Social Science Computer Review*, no. 23, vol. 4, 2005, pp. 486-501.
- [17] C. L. Borgman, "The conundrum of sharing research data.", in *Journal of the American Society for Information Science and Technology*, vol. 63, no. 6, 2012, pp. 1059-1078.
- [18] J. M. Wicherts, M. Bakker, D. Molenaar, "Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results", in *PLoS one*, vol. 6, no. 11, 2011, p. e26828.
- [19] H. A. Piwowar, R. S. Day, D. B. Fridsma, "Sharing detailed research data is associated with increased citation rate", in *PLoS one*, vol. 2, no. 3, 2007, p. e308.
- [20] P. Arzberger, P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, P. Wouters, "Promoting access to public research data for scientific, economic, and social development.", in *Data Science Journal*, vol. 3, 2004, pp. 135-152.
- [21] P. Langat, D. Pisartchik, D. Silva, C. Bernard, K. Olsen, M. Smith, R. Upshur, "Is there a duty to share? Ethics of sharing research data in the context of public health emergencies", in *Public Health Ethics*, vol. 4, no. 1, 2011, pp. 4-11.
- [22] A. L. McGuire, J. M. Oliver, M. J. Slashinski, J. L. Graves, T. Wang, P. A. Kelly, D. Treadwell-Deering, "To share or not to share: a randomized trial of consent for data sharing in genome research.", in *Genetics in Medicine*, vol. 13, no. 11, 2011, pp. 948-955.
- [23] J. C. Wallis, E. Rolando, C. L. Borgman, "If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology.", in *PLoS one*, vol. 8, no. 7, 2013, p. e67332.
- [24] M. Walport, P. Brest, "Sharing research data to improve public health", in *The Lancet*, vol. 377, no. 9765, 2011, pp. 537-539.
- [25] T. Hyla, J. Pejaš, "A practical certificate and identity based encryption scheme and related security architecture.", *Proceedings of 12th IFIP TC8 International Conference CISIM 2013, Krakow, Poland, LNCS vol. 8104, 2013, pp. 190-205.*
- [26] D. Pons, R. Vallés, M. Abarca, F. Rubio, "QR codes in use: the experience at the UPV Library." in *Serials [on-line]*, vol. 24, no. 3, Retrieved from <http://eprints.rclis.org/18047/1/QR%20codes%20in%20use.pdf>, 2011
- [27] W. Chmielarz, "Metody oceny werbalnych księgarni internetowych Komputerowo Zintegrowane Zarządzanie obszar: Gospodarka oparta na wiedzy", 2010, in Poland
- [28] W. Chmielarz, "Ocena użyteczności internetowych witryn sklepów komputerowych" *Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedzą Tom 13, 2008 s. 17-24*, in Poland
- [29] W. Chmielarz, „Przełączniki metodyczne w ocenie witryn internetowych sklepów komputerowych, Zarządzanie Wiedzą i Technologiami informatycznymi” red. C Ormowski, Z. Kowalczyk, E Szczerbinki, nr.4 seria: Automatyka i Informatyka, Pomorskie wydawnictwo Naukowo-Techniczne PWNT Gdańsk, v 43 2008, p. 361-368, in Poland
- [30] M. Socorro Garcia-Cascales, M. Teresa Lamata, Solving a decision problem with linguistic information. *Pattern Recognition Letters*, October 2011, Volume 22, Issue 5, pp 779-788
- [31] K.M.A. Al Harbi, Application of the AHP in project management. *International Journal of Project Management* 19(1): 19-27, 2001
- [32] T.L. Saaty, *The Analytic Hierarchy and Analytic Network Processes for the Measurement of Intangible Criteria and for Decision-Making*, w: *Multiple criteria decision analysis: State of the art surveys*, J. Figueira, S. Greco, M. Ehrgott, Springer, 2005, p. 345-405
- [33] T.L Saaty, How to make a decision: the analytic hierarchy process. *European Journal of Operational Research*, 48, 1990, p. 9-26
- [34] L. Thomas, T.L. Saaty. How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, 48:9-26, September, 1990
- [35] K. Muszyńska, J. Swacha, A. Miluniec, Z. Drajek, „Evaluation of eGuides: a discussion of approaches.” in *Information Management*, 2014, pp. 45-54