Journal of Social and Economic Statistics

Vol. 2, No.2, Winter 2013© 2012 Published by JSES.

# DATA QUALITY MEASUREMENT PRINCIPLES AND DIMENSIONS

**Daina ŠĶILTERE[a], Svetlana JESIĻEVSKA[b]**

## Abstract

*The quality of statistical data is essential for effective decision-making. The problem of evaluating the quality of statistical data is not a new paradigm as rapid methodological changes and globalization complicate the generation of high quality statistical data in all areas. The problem is on selecting appropriate criteria to evaluate the quality of statistical data, not just related to the intention of statistical survey, but also to the beliefs held by both statisticians and respondents. As a result there is a strong need to discuss this topic. In this paper authors provide a multi-dimensional approach of measuring the quality of official statistical data and propose the system of characteristics to determine statistical data quality. To some of the proposed data quality characteristics not much attention has been paid previously.*

**Key words:** statistical data, quality, measurement, validity, dimensions, quality measures
**JEL Classification**: C40

## Authors' Affiliation

[a]Dr.oec., Professor, Head of the Chair of economic systems management theory and methods, Faculty of Economics and Management, University of Latvia
[b]Mg.oec., Faculty of Economics and Management, University of Latvia, Central statistical bureau of Latvia, mozir@inbox.lv

## 1. Introduction

Since all types of research must respond to the agreed canons of quality (Marshall and Rossman 2006), we cannot avoid discussing them, in spite of their philosophical and practical complexity as well as the difficulty in defining what quality actually means or covers. Nowadays there are multiple different ways to define data quality and there is currently no commonly agreed definition on what data quality is. Different analysts and different agencies provide different answers (Brackstone 1999, Carson 2000, Pipino et al. 2002), but all agree that “data quality” is a multidimensional concept.

In this paper authors provide multi-dimensional approach of measuring the quality of official statistical data and propose the system of characteristics to determine statistical data quality.

## 2. Data quality overview

Confidence in the quality of the statistical data is a survival issue for a statistical office. If its information becomes distrustful, the reputation of the statistical office is called into question. But quality is not an easily defined concept, and has become an over-used term in recent years.

There are multiple different ways to define data quality and there is until now there is no commonly agreed definition on what data quality is. As an example, Wang and Strong (Wang and Strong,1996) define that qualitative data should fit for use by data consumers. Kahn, Strong, and Wang (Kahn, Strong, and Wang,2002) give define data quality as "conformance to specifications" and "meeting or exceeding consumer expectations". Redman (Redman 2001) suggests that data of high quality should fit for their intended uses in operations, decision making, and planning. One more aspect here is that data are free of defects and possess desired features.A popular definition for quality is fitness for use provided by Juran (Juran, 1974). Therefore, the interpretation of the quality of some data item depends on the needs of data users and the tasks this statistical data should serve. While one user may consider the data quality sufficient for a given task, it may not be sufficient for another task or another data user.

One positive aspect in the problem of defining data quality is that we recognize its importance (Dörnyei, 2007), at the same time unfortunately there is no guideline to a universally accepted convention in judging quality (Denscombe, 2003). In fact, there are various very general dimensions of data quality. These dimensions define the characteristics of data in measurable forms. A data quality dimension is defined as a set of data quality attributes that most data consumers react to in a pretty consistent manner (Wang, Ziad and Lee, 2001). The most commonly mentioned data quality characteristics from scientific literature summarized by authors are the following (see Table 1):

## Table 1.Data Quality characteristics

| Data quality characteristics | Definitions from the literature |
|---|---|
| Accessibility | Lee et al. defined as the ease and breadth of access to information (Lee et al. 2001). |
| Accuracy | Blackstone commented that the accuracy of statistical data requires that it is accessible, interpretable, coherent (Blackstone 2001). |
| Applicability | As Sandelowski (Sandelowski 1986) explained, generalization is a very broad concept as every research situation is made up of a particular researcher in a particular interaction with particular informants.<br>Guba (Guba 1981) refers to fittingness, or transferability, as the criterion against which applicability of qualitative data is assessed. |
| Authenticity | Authenticity deals with an obligation to improve the respondents' abilities to experience, understand andact in their reality. |
| Coherence | According to Vaismoradi and Salsali coherence describes the fit between the aim, the philosophical perspective, the researcher role in the study and the methods of investigation, analysis and evaluation undertaken by the researcher (Vaismoradi and Salsali 2010). |
| Confirmability | Confirmability means that conclusions, interpretations and recommendations should be traced back to their sources (Erlandson et al. 1993). |
| Credibility | Credibility deals with the focus of the research and means the level of confidence in how well data and processes of analysis address the intended focus (Polit and Hungler 1999).<br>According to Cornick, credibility relates to the degree to which data can be believed based on the ability of the researcher (Cornick 2006). |
| Neutrality | Neutrality is the freedom from bias in the research procedures and results (Sandelowski 1986).<br>Guba (Guba 1981) defines neutrality not as researcher objectivity but as data and interpretational confirmability. |
| Objectivity | Being objective means not to be influenced by personal feelings or opinion and not to be dependent on the mind for existence (Soanes and Stevenson 2003). |
| Reflexivity | Reflexivity means the capacity to reflect upon one's actions and values when producing data (Seale 1998, Gouldner 1972). |
| Relevance | Relevance is a key dimension as if the data does not address data users' needs and when the data user will find the data inadequate. |
| Reliability | Reliability means that data should be free from sources of measurement error and consistent (Creswell 2002). |
| Rigor | Rigor means that data is strict and inflexible (McKean 2005). |
| Security | Security means keeping data secure and restricting access to it. |
| Timeliness | Timeliness refers to whether data is current. |
| Transferability | Transferability means whether data can be used within other similar contexts (Houghton et al. 2012). |
| Trustworthiness | Trustworthiness is closely connected with validity and reliability (Seale 1999). Trustworthiness also includes the question of transferability (Polit and Hungler 1999).<br>Trustworthiness is composed of credibility, dependability, confirmability and transferability (Politet al. 2001). |
| Validity | Validityrefers to whether to whether measuring instrument is measuring what it was intended (Everitt 2002, p.388). |

Some quality characteristics like objectivity, security, confirmability, coherence, rigor, neutrality are not so commonly mentioned and defined. At the same time, accessibility, timeliness, accuracy, validity, reflexivity, credibility are widely discussed in the context of data quality.

Some more data quality criteria from the literature are the following: transferability, generalizability, ontological authenticity, reciprocity, dependability, fittingness, vitality, sacredness, goodness (Creswell 2002, Patton 2002, Spencer et al. 2003); fairness (Lincoln and Guba 2000); breadth and depth (Flick 1992); consensus, instrumental utility (Eisner 1991); openness and clarity (Cohen and Crabtree 2008); verisimilitude, integrity, verite' (Garman 1994); resonance (Tracy 2010); extrapolation, reciprocity, empathic neutrality (Patton 2002); locatability (Goodhue 1995); portability (Caby et al. 1995); appearance, comparability, precision, relevance, redundancy, context, informativeness, conciseness, importance, sufficiency, usefulness (Delone et al. 1992).

Authors made the conclusion that there are many different views on defining and determining data quality and no systematic approach. That is why in the next chapter authors providing their systematic approach of determining data quality.

## 3. The new system of data quality measurement

Before giving the systematic approach of determining statistical data quality authors make a distinction between three different phenomena that is data, information and knowledge. What is the difference between data, information and knowledge?

Checkland and Howell (Checkland and Howell,1998) suggest that information is structured data that has contextual meaning. Information becomes knowledge at the moment of its interpretation (Miller 2002). Nonakaand Takeuchi (Nonaka and Takeuchi 1995) understand information as a flow of messages, while knowledge is created by that very flow of information anchored in the beliefs and commitment of its holder. As a result knowledge is substantially related to human action. Sutter suggests that good quality information should satisfy criteria specified by the information user, together with a certain standard of requirement which depends to the use that is made of it (Sutter 1993).

The authors provide the following view of the link between data, information and knowledge that is based on the definitions mentioned above (see Fig.1).

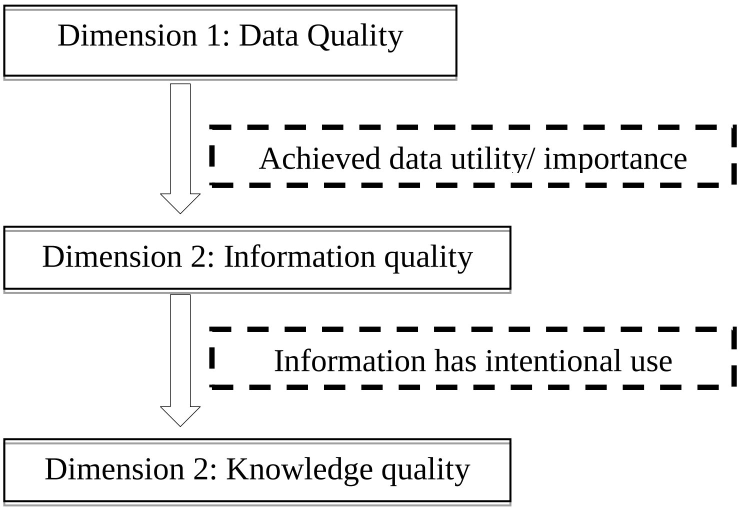
Dimension 1: Data Quality
Achieved data utility/ importance
Dimension 2: Information quality
Information has intentional use
Dimension 2: Knowledge quality

**Fig.1. Link between data, information and knowledge quality**

This study deals with the first dimension: data quality. Authors propose the following systematic approach for determining data quality that contains eleven characteristics of quality:

1. Validity
   - reliability
   - accuracy
   - representativeness
   - adequacy and substantiated nature of a measuring instrument
   - objectivity
2. Comparability
3. Completeness
4. Coherence
5. Understandability/interpretability/clarity of the data
6. Complexity
7. Flexibility
8. Timeliness/actuality in disseminating results
9. Utility/importance
10. Informativeness
11. Sensitivity.

These characteristics of data quality can be classified into 4 Dimensions (see Table 2).

**Table 2. Dimensional classification of data quality characteristics (authors')**

| **Dimensions** | **Data quality characteristics** |
|---|---|
| Dimension 1: Data users related quality characteristics | Understandability/interpretability/clarity of the data<br>Timeliness/actuality in disseminating results<br>Informativeness |
| Dimension 2: Data reporting and access related quality characteristics | Complexity<br>Comparability<br>Completeness<br>Flexibility |
| Dimension 3: Statistical process related quality characteristics | Validity<br>Coherence |
| Dimension 4: Institutional quality characteristics | Sensitivity<br>Utility/importance |

The proposed components of the data quality measuring system are defined and understood in the following way:

### 1) Validity

In the scientific literature validity is a term which can be applied to a lot of phenomenon; it can apply to a complete study and even to a whole theory and all its related empirical investigations. Brinberg and McGrath state that validity is like integrity, character, and quality, to be assessed in relation to aims and circumstances (Brinberg and McGrath 1985). Maxwell gives the similar idea that different methods can produce valid data in some circumstances and invalid ones in others (Maxwell 1992).As a result, the exact nature of 'validity' is a highly debated topic in the context of research since there is no single or agreed definition of this term.

Authors propose the following definition of validity: **The "validity" in the context of statistical data quality is in correspondence with reality that is supported by the adequacy and substantiated nature of a measuring instrument. Validity implies that statistical data should be accurately estimated and as a result data are of high validity if they are reliable, representative and objective.**Validity is a multi-dimensional and consists of the following characteristics: reliability, accuracy of estimates, representativeness, adequacy and substantiated nature of a measuring instrument, objectivity.

- Reliability means the closeness of the initial estimated value to the subsequent estimated value. Reliability involves comparing estimates over time or in other words, reliability refers to revisions. Generally speaking, the smaller and fewer the revisions, the better.
- Accuracy of estimates refers to the closeness between the estimated value and the true value that the statisticians measured. In practice, there is no overall measure of accuracy. Assessing the accuracy of estimates involves evaluating the error associated with an estimate.
- Representativity is often used in survey research, but usually it is not clear what it means. Kruskal and Mosteller (Kruskal and Mosteller1979a, 1979b and 1979c) found the following meanings for 'representative sampling': (1) general acclaim for data, (2) absence of selective forces, (3) miniature of the population, (4) typical or ideal case(s), (5) coverage of the population, (6) a vague term, to be made precise, (7) representative sampling as a specific sampling method, (8) as permitting good estimation, or (9) good enough for a particular purpose. The authors agree to this approach.
- Adequacy and substantiated nature of a measuring instrument means the correct methodology and the correct use of methodology.
- Objectivity can be understood simply as accurate, reliable, and unbiased information (Noe et al. 2003). In similar words this means whether the information was objectively collected.

2) **Comparability of statistics** refers to the degree to which statistical data are comparable over space (between countries) and time (between different time periods) as well as whether enough information is given to users to prevent any confusion when comparing statistical data.
3) **Completeness** means that statistical data should serve user needs as completely as possible, taking restricted resources into account.
4) **Coherence** between statistical data is orientated towards the comparison of different statistical data, which are produced in different way and for different primary uses. Coherence should be analyzed in the following aspects: data produced at different frequencies; other statistics in the same domain; sources and outputs; coverage of different databases; and definitions and coding used for different databases.
5) **Understandability/interpretability/clarity of the information** reflects the ease with which the user may understand and properly use and analyze the data. The adequacy of the definitions of concepts, target populations, variables and terminology underlying the

data largely determines their degree of understandability/interpretability/clarity of the information.

6) **Complexity** shows the possible difficulties that are connected with the processing of statistical data, usually expressed in the terms of resource consumption.
7) **Flexibility** refers to the data ability to be adjustable to the unique needs and to the rapidly changing environment.
8) **Timeliness/actuality in disseminating results** reflects the length of time between its availability and the event or phenomenon it describes, but considered in the context of the time period that permits the information to be of value and still acted upon.
9) **Informativeness** is a user-centred concept for evaluating the effectiveness of a statistical data. Informativeness indicates the raw potential a data has of informing a data user. Informativeness is particularly valuable due to its flexibility.
10) **Utility/importance** is the extent to which the statistical data compiled and supplied by the statistical agency is relevant to users' needs. In assessing the degree of utility/importance, three factors are taken into account: the analysis of main users; users' requirements, as identified by the statistical agency; and the level of users' satisfaction with the statistical information. The main difficulties in assessing relevance come from the fact that it is not easy to find out exactly who are the main users of certain statistical data and that the users' requirements may vary with time.
11) **Sensitivity** is confidence in the quality of the data and is a survival issue for a statistical agency. If its information becomes suspect, the reputation and the credibility of the agency is called into question.

## 4. Conclusions

In this article authors make an overview of existing theory on data quality issue. The fact that there are so many possible definitions for the term “data quality” and plenty of mentioned data quality indicators suggests that it is a common concept relative to the researcher and belief system for which it stems. Based on existing theory, authors developed a system of quality indicators to be used to determine the quality of statistical data. This systematic approach consists of the following quality characteristics: validity, comparability, completeness, coherence, understandability/interpretability/clarity of the data, complexity, flexibility, timeliness/actuality in disseminating results, utility/importance, informativeness, sensitivity. To some of these characteristics not much attention has been paid previously, for example, complexity, flexibility, informativeness, sensitivity.

In the context of proposed data quality system, authors provide the following understanding of validity: The “validity” in the context of statistical data quality is in correspondence with reality that is supported by the adequacy and substantiated nature of a measuring instrument. Validity implies that statistical data should be accurately estimated and as a result data are of high validity if they are reliable, representative and objective.

## References

Blackstone, G. (2001).*Managing data quality: the accuracy dimension*.The International Conference on Quality in Official Statistics, Stockholm, Sweden.

Brackstone, G. (1999).*Managing Data Quality in a Statistical Agency*. Survey Methodology, 25, pp. 139-149.

Brinberg, D. and McGrath, J.E. (1985).*Validity and the research process*. Newbury park, CA: Sage.

Caby, B.C., Pautke, R.W. and Redman, T.C.(1995).Strategies for improving data quality.*Data Quality*,1(1), 4-12.

Carson, C. S. (2000). *What is Data Quality? A Distillation of Experience*.Statistics Department, International Monetary Fund.

Checkland, P. and Howell, S. (1998).*Information, Systems, and Information Systems – Making Sense of the Field*.Chishester: John Wiley and Sons.

Cohen, D. J. and Crabtree, B. F. (2008). Evaluative criteria for qualitative research in health care: Controversies and recommendations. *Annals of Family Medicine*, 6(4), 331-339.

Cornick, P. (2006). Nitric oxide education survey – use of a Delphi survey to produce guidelines for training neonatal nurses to work with inhaled nitric oxide. *J. Neonatal Nurs*, 12(2), 62-68

Creswell, J. (2002). *Educational research: Planning, conducting And evaluating quantitative and qualitative research.* New Jersey: Pearson Education.

Delone, W.H. and McLean, E.R. (1992). Information systems success: The quest for the dependent variable.*Information Systems Research*,3(1), 60-95.

Denscombe, M. (2003).*The Good Research Guide (2 ed.).* Berkshire: Open University Press.

Dörnyei, Z. (2007). *Research methods in Applied Linguistics.* Oxford: Oxford University Press.

Eisner, E. W. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice.* New York: Macmillan Publishing Company.

Erlandson,D.A.,Harris,E.L.,Skipper,B.L. and Allen,S.D. (1993). *Doing Naturalistic Inquiry.A Guide to Methods*.Newbury Park: Sage.

Everitt, B.S. (2002).*The Cambridge Dictionary of Statistics Second Edition*, Cambridge Univesity Press.

Flick, U. (1992). Triangulation revisited: Strategy of validation or alternative? *Journal for the Theory of Social Behaviour*, 22(2), pp. 175-197.

Garman, N. (1994). *Qualitative inquiry: Meaning and menace for educational researchers (Keynote address).* Paper presented at the Mini-Conference: Qualitative Approaches in Educational Research, The Flinders University of South Australia.

Goodhue, D.L. (1995).Understanding user evaluations of information systems.*Management Science*.41(12), 1827-1844.

Gouldner, A.W. (1972).Towards a Reflexive Sociology.*In*: C. Seale *Social Resarch Methods*. London, Routledge, 381-383.

Guba, E. G. (1981).*Criteria for assessing the trustworthiness of naturalistic inquiries*.Educational Resources Infonnation Center Annual Review Paper, 29, 75-91.

Houghton, C., Casey, D., Shaw, D., Murphy, K. (2012). Rigour in qualitative casestudy research.*Nurse Res*, 20(4), 12-7.

J. Juran. (1974). *The Quality Control Handbook*. McGraw-Hill, New York, 3rd edition.

Kruskal, W. and Mosteller, F. (1979a). Representative sampling, I: Nonscientific literature. *International Statistical Review*, 47, 13–24.

Kruskal, W. and Mosteller, F. (1979b). Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review*, 47, 113–127.

Kruskal, W. and Mosteller, F. (1979c). Representative sampling, III: the current statistical literature. *International Statistical Review*, 47, 245–265

Lee, Y.W., Strong, D.M., Kahn, B.K., and Wang, R.Y. (2002). AIMQ: A methodology for information quality assessment. *Journal of Information and Management*, 40, 133-146.

Lincoln, Y.S., Guba, E.(2000). Paradigmatic controversies, contradictions and emerging confluences.*In*: Denzin, N.K., Lincoln, Y.S. (Eds.), *The Handbook of Qualitative Research*, Second ed. Sage Publications, Thousand Oaks, CA, 163–188.

Marshall, C. and Rossman, G.B. (2006).*Designing Qualitative Research (4 ed.).* Thousand Oaks, CA: Sage.

Maxwell, J. A. (1992).Understanding and validity in qualitative research.*In*: A. M. Huberman& M. B. Miles (Eds.), *The qualitative researcher's companion*, pp. 37-64. Thousands Oaks, CA: Sage Publications (Reprinted from Harvard Education al Review. 1992, 62, 3; 279-300).

McKean, Erin (Ed.). (2005). *The New Oxford American Dictionary (2$^{nd}$ ed.).* Oxford: Oxford University Press.

Miller, F.J.(2002). I=0 (Information has no meaning),*Information Research*, 8(1).

Noaka, I. and Takeuchi, H.(1995).*The Knowledge - Creating Company: how Japanese companies create the Dynamics of innovation.*New York: Oxford University Press.

Noe, P., Anderson, F., Shapiro, S., Tozzi, J., Hawkins, D., Wagner, W.(2003). Learning to live with the Data Quality Act.*Environ Law Rev*., 33, pp. 10224-10236.

Patton, M. Q. (2002). *Qualitative research and evaluation methods (3$^{rd}$ed.).*Thousand Oaks, California: Sage Publications.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002).Data Quality Assessment.*Communications of the ACM*, 45, pp. 211-218.

Polit, D.F. andHungler, B.P. (1999). *Nursing Research.Principles and Methods, sixth ed.* J.B. Lippincott Company, Philadelphia, New York, Baltimore.

Polit, D., Beck, C., Hungler, B. (2001).*Essentials of Nursing Research – Methods, Appraisal and Utilisation.*Philadelphia: Lippincot.

Redman, T. C. (2001). *Data quality: the field guide.* Boston: Digital Press.

Sandelowski, M. (1986).The problem of rigor in quaJitative research.*Advances in Nursing Science*,8, 27-37.

Seale C. (1998). *Researching Society and Culture*. London: Sage.

Soanes, C. and Stevenson, A. (2003).*Oxford dictionary of English*, Oxford, UK Oxford University Press.

Spencer, L., Ritchie, J., Lewis, J. and Dillon, L. (2003). Q*uality in qualitative evaluation: A framework for assessing research evidence.* London: National Centre for Social Research, Government Chief Social Researcher's Office, UK.

Sutter, E. (1993). Maîtriser l’information pour garantir la qualité. AFNOR.

Tracy, S. J. (2010). Qualitative quality: Eight “Big-Tent” criteria for excellent qualitative research. *Qualitative Inquiry*, 16(10), 837-851.

Vaismoradi, M. and Salsali, M. (2010). Coherence in qualitative research. *Journal of Nursing and Midwifery*, 20(70).

Wang, R. Y., Ziad, M.and Lee, Y. W. (2001). *Data quality*. Massachusetts: Kluwer Academic Publishers.

Wang, R.Y. and Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers.*Journal of management Information Systems*, 12(4), 5-33.