Analysis of Data Quality and Information Quality Problems in Digital Manufacturing

K. Q. Wang[1], S. R. Tong[1], L. Roucoules[2], B. Eynard[3]
[1]School of Management, Northwestern Polytechnical University, Xi'an, China
[2]Laboratory of Mechanical Systems and Concurrent Engineering, University of Technology of Troyes, Troyes, France
[3]Department of Mechanical Systems Engineering, University of Technology of Compiègne, Compiègne, France
(keqin.wang@nwpu.edu.cn, stong@nwpu.edu.cn, lionel.roucoules@utt.fr, benoit.eynard@utc.fr)

Abstract – **This work focuses on the increasing importance of data quality in organizations, especially in digital manufacturing companies. The paper firstly reviews related works in field of data quality, including definition, dimensions, measurement and assessment, and improvement of data quality. Then, by taking the digital manufacturing as research object, the different information roles, information manufacturing processes, influential factors of information quality, and the transformation levels and paths of the data/information quality in digital manufacturing companies are analyzed. Finally an approach for the diagnosis, control and improvement of data/information quality in digital manufacturing companies, which is the basis for further works, is proposed.**

Keywords – **Data quality, information quality, digital manufacturing**

Note: In data quality/information quality fields, unless specified otherwise, most papers use "information" interchangeably with "data". This paper will follow the same rule.

I. INTRODUCTION

In digital manufacturing (DM) industry, large amount of data and information has been collected during the product manufacturing process. However, much of the data/information has quality problems, and Data Quality /Information Quality (DQ/IQ) problems are becoming increasingly evident, particularly in manufacturing databases. Poor DQ/IQ in DM can severely hamper organizations' effectiveness and these problems are pervasive and costly [1][2]. As Davenport stated, "no one can deny that decisions made based on useless information have cost companies billions of dollars" [3]. Solving DQ problems typically requires a very large investment of time and energy - often 80% to 90% of a data analysis project is spent in making the data reliable enough that the results can be trusted [4].

There have been much works on DQ/IQ problems, however, most works focused on general database field and others focused on DQ in accounting, spatial, statistical, healthcare, environment fields, etc. Few works have been done on the analysis and solving of DQ problems in DM field.

For the purpose of the improvement of DQ/IQ in DM industry, this paper aims to analyze the DQ problems in DM companies. Particularly, some fundamental issues in this field will be investigated.

This paper is organized as follows: Section II reviews related works on DQ/IQ researches. Section III proposes the different information roles in DM, information manufacturing process, influential factors of IQ, and the transformation levels and paths of the DQ/IQ in DM companies. An approach for the diagnosis, control and improvement of DQ/IQ in DM is proposed in Section IV, which is the basis for further works. Finally, Section V concludes this paper with discussion and future works.

II. RELATED WORKS

This section will review some related works which cover the following four questions of DQ: (1) What is the definition of data quality? (2) What are the dimensions of data quality? (3) How is data quality measured and assessed? (4) How to deal with poor quality data?

A. DQ Definitions

The research group led by Professor Strong from MIT is one of the most successful groups in DQ field. Adopted from the definition of "quality" by Juran [33], Strong et al. defined DQ as *fitness for use* by data consumers [5], which is a widely adopted criterion.

From the standpoint of feedback-control system, DQ is actually quite easily defined as the measure of the agreement between the data views presented by an information system and that same data in the real world [1]. A system's DQ of 100% would indicate, for example, that our data views are in perfect agreement with the real world, whereas a DQ rating of 0% would indicate no agreement at all. Now, no serious information system has DQ of 100%. The real concern with DQ is to ensure that the DQ system is accurate enough, timely enough, and consistent enough for the organization to survive and make reasonable decisions.

B. DQ Dimensions

Just as a material product has quality dimensions associated with it, an information product has IQ dimensions [6]. Many scholars have proposed different numbers of DQ/IQ dimensions. Wang concluded that there was no general agreement on DQ/IQ dimensions [7]. There are three primary types of researches who have attempted to identify appropriate DQ dimensions: 1) data quality, 2) information systems, and 3) accounting and auditing.

In DQ area, Ballou et al. [9-12] defined four DQ

dimensions: 1) accuracy, which occurs when the recorded value is in conformity with the actual value, 2) timeliness, which occurs when the recorded value is not out of date, 3) completeness, which occurs when all values for a certain variable are recorded, and 4) consistency, which occurs when the representation of the data value is the same in all cases. Wang identified DQ/IQ with four DQ/IQ categories and fifteen dimensions [13], as shown in Table 1. Others identified DQ dimensions as data validation, availability, traceability, and credibility, and so on [14-16].

In the information systems area, Halloran et al. [17] proposed various factors such as usability, reliability, independence, etc. Kriebel [18] identified attributes as accuracy, timeliness, precision, reliability, completeness, and relevancy, Ahituv [19] suggested relevant attributes such as timeliness, accuracy, and reliability.

Many works in the accounting and auditing literature specifically emphasized on internal control systems and audits [34][35], where internal control systems require maximum reliability with minimum cost, the key DQ dimension used is accuracy - defined in terms of the frequency, size, and distribution of errors in data. Others, for example, Feltham [36] identified relevance, timeliness, and accuracy as the three dimensions of DQ.

C. DQ Measurement and Assessment

Commonly used methods for measurement of DQ and/or IQ are through multiple data dimensions. Recent years Wang and his team focus on Total DQ Management (TDQM) based on the Total Quality Management (TQM). The TDQM methodology adapted for the evaluation of DQ in an information system (by assuming that each piece of produced information can be considered as a product) [6]. Following the TQM cycle (Definition, Measurement, Analysis and Improvement), the measurements step produces the quality metrics. Lee et al. [20] developed the AIMQ (AIM Quality) methodology for assessing and benchmarking IQ in organizations, which has been applied in manufacturing industry.

Pierce assesses DQ with Control Matrices [21]. Cappiello proposed and verified one model for assessing DQ from the user's perspective [22]. Ref. [23] developed a quantitative measure of DQ by formulating the error rate of MIS records, which are classified as being either correct or erroneous. Ref. [24] showed how subjective quality goals were evaluated using more objective quality factors. In DaQuinCIS system [25][26], data source providers were evaluated by data source users in a peer-to-peer system. Unfortunately, the system relied heavily on the participation of users in the review of the quality of data in the system, which might not be practical. Ref. [27] proposed detailed IQ evaluating indicators and evaluated the IQ by AHP method. Zhang [28] and Su et al. [29] studied much about manufacturing information TQM. They evaluated the quality of manufacturing information through five quality variables such as functionality, dependability, timeliness, usability, and economy. Then the manufacturing information could be evaluated at length by three quality variable sets which include 30 quality variables in total which belonged to the five aspects mentioned above.

TABLE I
DQ CATEGORIES AND DIMENSIONS [13]

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

D. DQ Improvement

Conventional approaches employ control techniques (like edit checks, database integrity constraints) to ensure DQ. The approaches have improved intrinsic DQ substantially, especially the accuracy dimension. However, attention to accuracy alone does not correspond to data consumers' broader DQ concerns. Furthermore, controls on data storage are necessary but not sufficient [5].

In the TQM cycle of TDQM methodology, the "improvement" step provided techniques for improving IQ [6]. The AIMQ methodology [20] is useful in identifying IQ problems, prioritizing areas for IQ improvement, and monitoring IQ improvements over time.

Winkler [30] proposed methods for evaluating and creating DQ. The author presented a statistician perspective on methods for statistical data editing and imputation and for data cleaning to remove duplicates. Scannapieco et al. [26] proposed DaQuinCIS architecture which is a platform for exchanging and improving DQ in cooperative information systems. Ref. [11][31] presented various analytical models and procedures for data enhancement in database and data warehouse environments. Helfert, Zellner, and Sousa [32] proposed some means ensuring DQ. Ken Orr [1] claimed that one certain way to improve the quality of data: improve its use!

III. MANUFACTURING PROCESS OF DQ/IQ IN DM

There are different roles in information related processes of DM. At the same time, the information will be regarded as product which is produced during the manufacturing processes. This section will analyze the information roles in the information manufacturing process (IMP). Then the transformation levels and paths of IQ will also be analyzed for further investigation of weak points along with the transformation paths.

A. Information Roles in DM

Everyone in DM companies has to use information. Thus, all the people act as different types of roles in the

information related processes. We adopt the perspective that the information roles include information provider, information processor, information manager, and information consumer.

The information roles can be discussed from three perspectives. Firstly, the same person or entity (i.e. other information processing units) can act as information provider, information processor, information manager, or information consumer. For example, the process designer is information consumer of the product parameter design information, at the same time he is information provider for manufacturing engineers, as illustrated in Fig. 1(a).

Secondly, the same information can be consumed by different people or entities. For example, as regarding to the same parameter design information, both the process designer and manufacturing engineers may be information consumer, as illustrated in Fig. 1(b).

Thirdly, different information consumer may require the same information. The same information consumer may require different information. As illustrated in Fig. 1(c).

B. Information Manufacturing Process

In DM companies, different types of information are manufactured, just like the manufacturing process of material product, we call them information products (IP). The information is roughly classified into three categories: product design information, production information, and management information.

The manufacturing process of each type of information starts from the information sources, along with gathering, processing, storing, and transformation, finally arriving at the information consumers. Besides above-mentioned activities included in the manufacturing process, some other activities may be always accompanied including maintenance, management, and

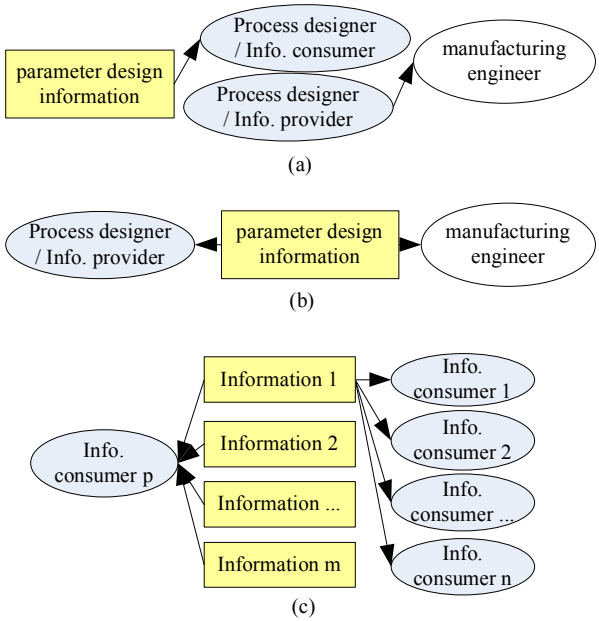

Fig. 1. Different information roles.

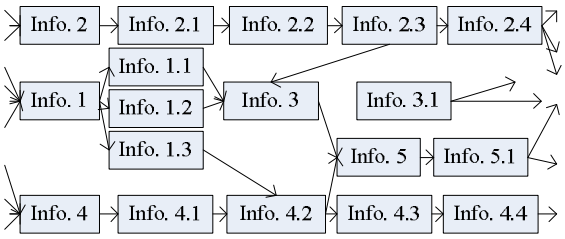

Fig. 2. Information manufacturing process.

updating of the information.

In IMP of IP, the information undergoes complicated changes. Some typical changes of the IP are described as follows: Firstly, one piece of information may diverge into different pieces of information for different next-step information processor. Secondly, on the contrary, many different pieces of information may become converged into one piece of information for further consumption. Thirdly, the same information may be just processed without any interaction with other pieces of information. Here we call it the information serialization. Fourthly, different pieces of information may never be converged into one piece of information, called parallel information.

In fact, most of the information interaction in the IMP includes the four above-mentioned types of information changes. The four types of information changes are described in Fig. 2. Here the information processors are ignored deliberately in order to see the information changes clearly.

C. Influential Factors of Information Quality

There exist many types of IQ problems in DM companies. All the IQ problems may be influenced by some specific factors. As we know in quality management field, the quality problems are often analyzed through 5M1E (Man, Machine, Material, Methods, Measurement and Environment). Just as we call IP as well as material product, the IQ problems in DM companies can also be analyzed through 5M1E. This paper will analyze the major influential factors of IQ problems along with 5M1E.

Here we propose the meaning of 5M1E factors which influence the IQ problems in DM. Men, the people who act as different information roles in the DM, of course have influence on the IQ during the IMP. Machine, in IQ problems means information processing units such as database, information systems, etc. Material is raw data or raw information for further processing. Methods mean different approaches on how information roles process information. Measurement plays important roles in IQ assessment and evaluation. Different kind of measurement may result in different IQ precision and cause different IQ problems. The 5M1E factors will be presented next for the analysis of the transformation levels and paths of IQ.

D. Transformation Level and Path of Information Quality

Along with the IMP, the IQ is also transformed at the

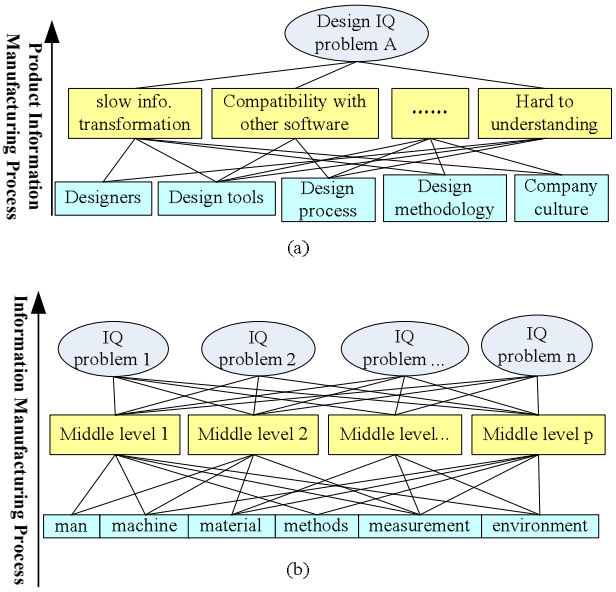

Fig. 3. Transformation level and path of IQ.

same time. Meanwhile the transformation path of IQ can also be identified. We call it the information transformation levels and paths based on IMP. As mentioned above, there are 5M1E factors influencing the IQ and may cause IQ problems. Thus it is necessary to present the transformation levels and paths along with the IQ manufacturing process.

An example is illustrated in Fig. 3 for understanding of the transformation level and path of IQ. As shown in Fig. 3(a), the design IQ problem A can be analyzed by decomposition into next level until the complete IQ transformation level and path. Fig. 3(b) shows the whole hierarchical model of the different IQ problems and its influential factors.

What should be noted is that the situations illustrated in the figures are just for purpose of analysis. In real DM companies, the situation may be much more complicated.

IV. APPROACHES TO IMPROVE DQ IN DM

Based on the analysis of information roles, IMP, the influential factors of IQ problems, and the transformation levels and paths of IQ, the diagnosis, control and improvement of the IQ level for DM companies are the final goals of our project. The IQ project undertaken by our team proposes one approach for the diagnosis, control and improvement of IQ, as shown in Fig. 4.

Note that the detail operation is not as simple as illustrated in the figure. Fig. 4 is just for purpose of illustration of our approach. The detailed content of the approach will be discussed in further works. The IQ diagnosis and control will adopt the SPC (Statistical Process Control) toolkit and Six Sigma methodology. The SPC toolkit includes historical diagram, Pareto diagram, fishbone diagram, control chart etc [37]. The six sigma

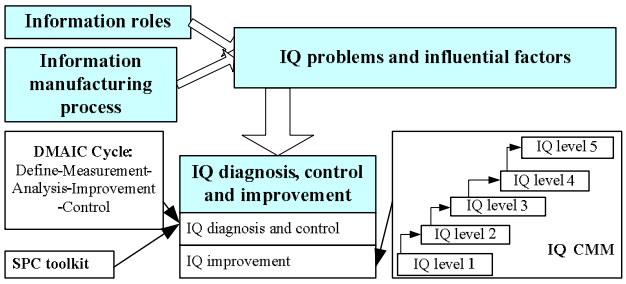

Fig. 4. Approach for IQ diagnosis, control and improvement.

methodology [38] is about the DMAIC cycle (Define – Measurement – Analysis - Improvement - Control).

The improvement of IQ level in DM companies will adopt the CMM (Capacity Maturity Model) approach which is popular in software engineering field. There are five levels in the CMM model, level 1 is the lowest and level 5 is the highest. Most companies are in the level 2 or 3. It is hard for most companies to go up to level 5. The DM companies can identify its IQ level through the analysis of its IQ situation, and then propose the improvement goal of next operation. The detailed standard for the five levels in DM IQ situation will be defined in future works.

V. CONCLUSIONS AND FUTURE WORKS

DQ/IQ problems are becoming increasingly evident in DM companies. It is clear that wrong data is likely to result in wrong decisions in manufacturing process. The literature review shows that few DQ/IQ works has been done in DM fields even we have already recognized the importance of DQ/IQ in DM. By the analysis of the information roles, IMP, influential factors of IQ problems, and the transformation levels and paths of IQ in DM, it will be clear to know where the IQ weak points may exist. The relationships identified between DQ/IQ and its influential factors are valuable for manufacturers to investigate and solve the DQ/IQ problems. The approach proposed for the diagnosis, control and improvement of DQ/IQ in DM is the final goal of our project.

However, this paper just analyzed some fundamental issues concerning the DQ/IQ problems in DM. Some works need to be done in future. The IQ problems must belong to different modes, how to identify these IQ problem modes is important. The relationship model between the IQ problems and their influential factors need to be investigated in detail. How to diagnose and control the DQ/IQ is the core work for manufacturers. The DQ maturity model should be built for the evaluation of the DQ/IQ level of the manufacturing companies.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of National Natural Science Foundation of China (NSFC,

No. 70771091), the Aeronautics Science Foundation of China (No. 2007ZG53074), and the Youth for NPU teachers Scientific and Technological Innovation Foundation.

REFERENCES

[1] K. Orr, "Data quality and system theory," *Comm. ACM*, vol. 41, no. 2, pp. 66-71, Feb. 1998.

[2] T. C. Redman, "The impact of poor data quality on the typical enterprises," *Comm. ACM*, vol. 41, no. 2, pp. 66-71, Feb. 1998.

[3] T. H. Davenport, *Information Ecology: Mastering the Information and Knowledge Environment*, New York: Oxford University Press, 1997.

[4] T. Johnson and T. Dasu, "Data quality and data cleaning: an overview," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, San Diego, CA, pp. 681-681, 2003.

[5] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," *Comm. ACM*, vol. 40, no. 5, pp. 103-110, May 1997.

[6] R. Y. Wang, "A product perspective on total data quality management," *Comm. ACM*, vol. 41, no. 2, pp. 58-65, Feb. 1998.

[7] R. Y. Wang, V. C. Storey, and C. P. Firth, "A framework for analysis of data quality research," *IEEE Trans. Know. Data Eng.*, vol. 7, no. 4, pp. 623–640, 1995.

[8] D. P. Ballou and H. L. Pazer, "The impact of inspector fallibility on the inspection policy serial production system," *Management Science*, vol. 28, no. 4, pp. 387-399, 1982.

[9] D. P. Ballou and H. L. Pazer, "Modeling data and process quality multi-input, multi-output information systems," *Management Science*, vol. 31, no. 2, pp. 150-162, 1985.

[10] D. P. Ballou and H. L. Pazer, "Cost/quality tradeoffs for control procedures information systems," *OMEGA: Int'l J. Management Science*, vol. 15, no. 6, pp. 509-521, 1987.

[11] D. P. BaJlou and K. G. Tayi, "Methodology for allocating resources for data quality enhancement," *Comm. ACM*, vol. 32, no. 3, pp. 320-329, 1989.

[12] D. P. Ballou, R. Y. Wang, H. Pazer, and K. G. Tayi, "Modeling data manufacturing systems to determine data product quality," Report No. TDQM-93-09, Cambridge, Mass, 1993.

[13] R. Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, No. 4, pp. 5-34, Spring 1996.

[14] R. Y. Wang, M. P. Reddy, and A. Gupta, "An object-oriented implementation of quality data products," in *Proc. Third Ann. Workshop Information Technologies and Systems*, Orlando, FLA, pp. 48-56, 1993.

[15] M. Janson, "Data quality: the achilles heel of end-user computing," *Omega J. Managemenr Science*, vol. 16, no. 5, pp. 491-502, 1988.

[16] G. E. Liepins and V. R. R. Uppuhui, "Accuracy and relevance and the quality of data," in *Data Quality Control Theory and Pragmatics*, A. S. Loebel, Ed., New York: Marcel Dekker, 1990, pp. 105-143.

[17] D. Halloran et al., "Systems development quality control," *MIS Quarterly*, vol. 2, no. 4, pp. 1-12, 1978.

[18] C. H. Kriebel, "Evaluating the quality of information systems," in *Design, and Implementation of Computer Based Information Systems*, N. Szysperski and E. Grocbla, Eds. Germantown: Sijthtoff and Noordhoff, 1979.

[19] N. Ahituv, "A systematic approach toward assessing the value of an information system," *MIS Quarterly*, vol. 4, no. 4, pp. 61-75, 1980.

[20] Y. Lee, et al., "AIMQ: a methodology for information quality assessment," *Information and Management*, vol. 40, no. 2, pp. 133-146, 2002.

[21] E.M. Pierce, "Assessing data quality with control matrices," *Comm. ACM*, vol. 47, no. 2, pp. 82-86, Feb. 2004.

[22] C. Cappiello, C. Francalanci, and B. Pernici, "Data quality assessment from the user's perspective," in *Proceedings of IQIS 2004*, Maison de la Chimie, Paris, pp. 68-73, 2004.

[23] D. B. Paradice and W. L. Fuerst, "An MIS data quality methodology based on optimal error detection," *J. Information Systems*, vol. 5, no. 1, pp. 48-66, 1991.

[24] P. Vassiliadis, M. Bouzeghoub, and C. Quix, "Towards quality-oriented data warehouse usage and evolution," *Information Systems*, vol. 25, no. 2, pp. 89-115, 2000.

[25] M. Mecella, et al., "Managing data quality in cooperative information systems," in *CoopIS/DOA/ODBASE 2002, LNCS 2519*, R. Meersman and Z. Tari, Eds. pp. 486-502, 2002.

[26] M. Scannapieco, et al., "The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems," *Information Systems*, vol. 29, no. 7, pp. 551- 582, Oct. 2004.

[27] R. C. Cao and J. M. Wu, "Information quality and the evaluation index system" (in Chinese), *Information Research*, no. 4, pp. 6-9, Dec. 2004.

[28] B. P. Zhang, "Investigation on TQM for manufacturing information" (in Chinese), *Manufacturing Automation*, vol. 24, no. 8, pp. 1-5, 2002.

[29] Y. Su, M. Yu, and B. P. Zhang, "An approach for information quality assessment for mechanical product" (in Chinese), *China Mechanical Engineering*, vol. 15, no. 6, pp. 520-524, March 2004.

[30] W. E. Winkler, "Methods for evaluating and creating data quality," *Information Systems*, vol. 29, no. 7, pp. 531–550, Oct. 2004.

[31] D. P. Ballou and G. K. Tayi, "Enhancing data quality in data warehouse environments," *Comm. ACM*, vol. 42, no. 1, pp. 73-78, 1999.

[32] M. Helfert, G. Zellner, and C. Sousa, "Data quality problems and proactive data quality management in data warehouse systems," in *Proceedings of BITWorld*, Guayaquil, Ecuador, 02-05 June 2002.

[33] J. M. Juran and A. B. Godfrey, *Juran's Quality Handbook (5th Edition)*. New York: McGraw-Hill, 1999.

[34] D. Kaplan, R. Krishnan, R. Padman, and J. Peters, "Assessing data quality in accounting information system," *Comm. ACM*, vol. 41, no. 2, pp. 72-78, 1999.

[35] S. Divorski and M. A. Scheirer, "Improving data quality for performance measures: results from a GAO study of verification and validation," *Evaluation and Program Planning*, vol. 24, no.1, pp. 83-94, 2001.

[36] G. Feltham, "The value of information," *Accounting Rev.*, vol. 43, no. 4, pp. 684-696, 1968.

[37] J. S. Oakland, *Statistical Process Control (Sixth Edition)*. Oxford: Butterworth-Heinemann, 2007.

[38] T. Pyzdek, *The Six Sigma Handbook: The Complete Guide for Greenbelts, Blackbelts, and Managers at All Levels (2nd revised edition)*. New York: McGraw-Hill, 2003.