Data Quality Assessment Model for Improving Decision Making Process

Nasser Ahmed Hussein Al-Mughni[1], Abdullah Saad AL-Malaise AL-Ghamdi[2], Farrukh Saleem[3]

Information Systems Department,
Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah,
Kingdom of Saudi Arabia[1,2,3]

Abstract: Data in the organization creates a large impact on company's operational and business processes. It will help the organization to be consistent in fulfil the business context requirements, reduce the risk, smooth transactions of business processes, and improve the performance. Currently, the type and size of the data generating everyday requires proper techniques to maintain to integrate and maintain the quality of the data. Due to its importance and vast dimensions, maintaining the data quality is always being a big challenge for the organization. Furthermore, the data quality generates efficient results and state-of-the-art opportunities in different scenarios such as decision making process. Negligence in data collection and integration, inconsistent data quality attributes and poor systematic approach are major drawbacks of data quality. The use of simultaneous strategies between data collection, analysis and for decision making can help to optimize data quality and developing efficient business strategies. This research is focused on data quality assessment to improve the decision making process. Assessing data quality dimensions to improve the business performance and for efficient decision making process is the major concerned in this research. This research integrates five different dimensions including cognitive assessment, in proposed data quality framework. Whereas the implementation of the model on any industry will provide the reasonable understanding of the model to improve the data quality and decision making process.

Keywords: Data Quality, Data Quality Assessment, Decision Making, cognitive assessment

I. INTRODUCTION

Nowadays, organizations are more dependent on data where it has many uses such as creating new business strategies, performing routine functional processes, generating reports, and dealing with the customers. Apart from collecting and managing the data, the real challenge for the organization is how to maintain data and information quality for repetitive business processes [1]. Maintaining the data quality has large impact on organizational business operations, it allows the organization to fulfil customer's need, enhancing the process quality, help to predict more accurately, reduce the risk, and generating high quality data [2]. Data quality assessment for improving the decision making process is the main idea undertaken in this study. To more elaborate, the data plays significant role in generating multiple alternatives during decision making process. Specifically, the analysis and intelligent phase in that process are the key factors, which help to generate multiple alternatives for decisions [3]. In addition, the idea has been presented and validated several times in past that proves the association between data quality and decision making process [1], [4], [5]. The research will discuss the possible attributes for assessing the data quality, and how that assessment can enhance the decision performance.

Conceptually, data and information in the researches has been used for same purpose as defined in previous literature and explained by different scholars [6], [7]. Therefore, the list of attributes are also similar either using for data or information, but most of the time the term "Data" is commonly using in research and case studies [4], [8]. Data quality as discussed in the previous studies has focused on different perspective and presented as multi-dimensional concept [9]. The concept of multi-dimension is that quality cannot be achieved in single straight way but it requires multiple direction to be assessed. Therefore, the first systematic approaches which recognized and explained the data quality dimension [10], [11] presented using three different kinds of techniques to assess the data quality. Those three approaches are known as empirical, theoretical, and intuitive. The result presented in that study highlighted the importance of assessing the data quality from three different perspectives [12].

The next section discusses the further related studies to discuss the importance of this research. It also describes the common framework used for assessing the data quality. The list of attributes, types of data quality, most related work, and finally the problem statement presented in the subsequent section to give more connectivity in this research.

II. LITERATURE REVIEW

Maintaining data quality is a set of procedures in which implementation is performed on data contents using assessment techniques to generate valuable data with high quality and accuracy. Despite, the poor data quality can lead to unfortunate strategic decisions and false reports generation [13]. Improving data quality, collecting data, and utilization of data adequately are some challenges facing in past years for all kind of organization and businesses. Whereas decision making based on the high-quality data can give a competitive edge to any organization [14]. Keeping quality and proper processing of raw data can derive information, which can be used as intelligence in decision making process. Further different studies has described that the assessment of quality data can be applied using different attributes such as completeness, consistency, uniqueness, timeliness, validity, accuracy, and consistency [15]. Assurance of good quality data is becoming an integral component for quality decision making and management in any organization. Good quality data leads to the efficient decision making whereas poor quality data negatively influence the decision making and ultimately creates bad impact on organizational performance [16].

As per the above discussion, it is evident that data quality is an essential component and plays vital role in determining the decision making and organizational productivity [17]. At organizational level, the use of quality data is necessary for planning activities, improving productivity, strategic management and for operational processes [18]. However, the good quality data is being considered an essential prerequisite for the organization success. Due to the increasingly experienced drawbacks of low-quality data [19] the organizational efficiency and effectiveness can create negative impact [20]. Recent advances in technologies generating big amount of data every day, can cause substantial economic inefficiencies due to the low performance of data [21]. Generally, data quality is critical for determining the organizational and societal activities and outcomes. Moreover, the inadequate data results in increased time and cost spent on revising the ineffective management decisions to accomplish the desired organizational objectives and goals. Hence, the efficiency of decision making is related to the quality of data and organizational efficiencies can be improved with good quality data. [2] described the five major dimensions for data quality known as availability, usability, reliability, relevance and quality of presentation. Those elements identified for efficient practicing as shown in Fig. 1.
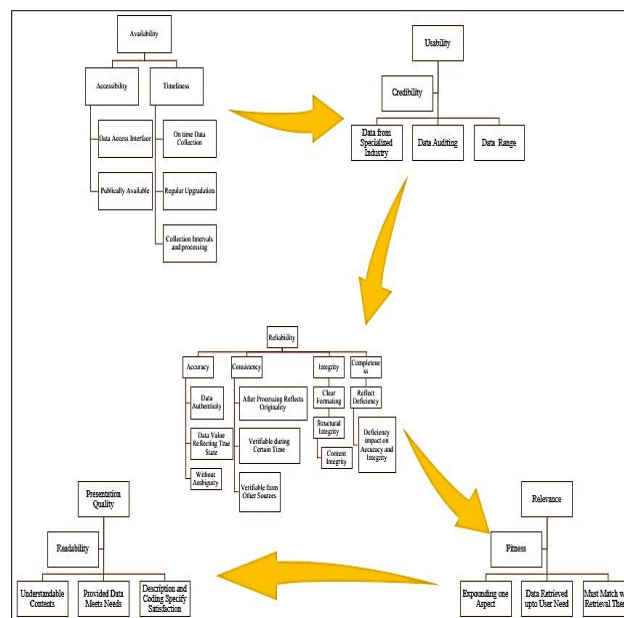

Fig. 1: The Hierarchy of Data Quality Assessment [2]

Data quality assessment greatly influence the quality of data and decision making. Data assessment involves the identification of deficiencies, error and control methods to improve the business strategies and implementation [9]. Whereas, the quality assessment of data involves the several kinds of methods such as qualitative and quantitative [22]. These data assessment involves the some other approaches as used and presented earlier [23], [24].

For the decision makers, it is important to understand the data analyses for business processes and requirements. In the organization, for every decision making process, enough data and its quality are essentially required. On the other hand, in the presence of good quality data other dimensions need to be considered as well. There are several data quality dimensions and attributes from researchers. Wang and strong proposed four different dimensions for data quality and

categorized them in further sub-dimensions [11]. In the same way, there are other researches that presented multiple factors for measuring the data quality [17], [25]. Further, the next section presents the detailed analysis about several related research and proposed a framework including the list of data quality attributes proposed.

III. PROPOSED FRAMEWORK

According to the findings from literature review, the summary of the variables and related work already presented in previous chapter. Based on those finding the model is proposed in this research as shown in Fig. 2. As discussed in the previous section, the model development in this research were based on the multiple models proposed for data quality assessment and its impact over decision making process.

Firstly, the data quality assessment framework presented by [11] where they proposed the model after several modification. The framework is applicable in different scenario, and has been used by different scholars. The framework is known as a pioneer in data quality assessment field. Therefore, the researcher selected this model to be applied in healthcare industry using Saudi case studies. As per the knowledge of the researcher and based on the online available literature, this model has not been applied on Saudi health industry yet. The main idea behind selection of this model is to improve the data quality in health industry using different tools. Whereas, the short listed data quality framework covers very large aspects of data quality using four different categories with fifteen sub-items.

Furthermore, the model has been integrated with the idea of cognitive assessment data quality approach proposed by [26] for covering the data quality improvement from cognitive perspective. The model was originally based on the idea to cover the user's contextual perspective on improving the data quality. Human cognition was the main idea in this research presented to improve the decision making process in healthcare industry. The author proposed this model to intervene the data quality by adding human cognition parameter based on their expertise, experiences, and understanding of the given tasks. Finally, the integrated framework connected with the decision making process. The idea is based on the research presented by [27] that shows how data quality can improve the overall process of decision making. The following is the integrated framework for data quality, which has been improved by adding cognitive assessment perspective. The model can support the organization by creating the knowledge base, which can be useful for future correspondence. Further, it can be directly connected with the decision maker for live interaction.
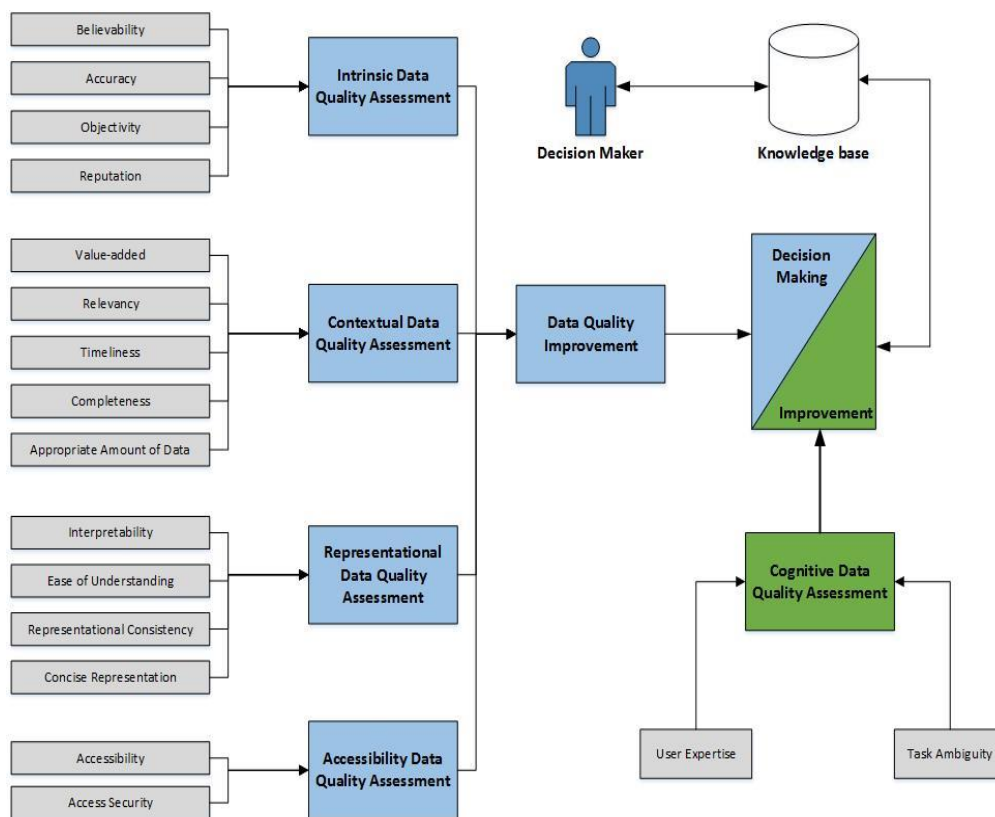

Fig. 2 Proposed Integrated Data Quality Framework for Improving Decision Making Process

A. Operational Definitions of Constructs used in Proposed Framework

The integrated conceptual framework presented in the previous section. The conceptual framework need to be operationalized by defining its major components and measuring item under each component. All constructs and measuring items also known as major variables that work for implementing the overall methodology of the research. There are different types of variables used in the research such as independent variable, dependent variables, and intermediate variables [28]. The purpose of defining the variable is to execute the research methodology and create output based on the values of each variable [29].

The independent variables are known as influence factor, which create the impact on other variables called dependent variable. Whereas, the mediating variables are those factors which support the research methodology by creating their impact between dependent and independent variables [28]. In other words, as defined by [30] that intermediate variables actually used to support the framework and alters the association between independent and dependent variables. Based on the understanding, this research used 41 measuring items used to assess the performance of 15 independent variables. In which 35 measuring items belong to the main dimensions of this study for measuring the data quality. Whereas, 6 measuring items are used to support the two intermediate variables selected for measuring the cognitive data quality assessment. As a whole there are five major dimensions proposed in the framework as extracted from literature review defined as intrinsic data quality, contextual data quality, representational data quality, accessibility data quality, and cognitive data quality. The following tables representing the operational definition of each variables used in this study.

The first major factor used in this research to measure the data quality in an organization. The factor known as "Intrinsic Data Quality", which has four different variables to assess this dimension from different perspective. The meaning of intrinsic is refer to the key areas of data quality. Therefore, the four constructs defined to measure this factor belong to the core area of data quality. Table I illustrates the definition of each construct.

TABLE I INTRINSIC DATA QUALITY - OPERATIONAL DEFINITION

Construct	Operational Definition	Source
Believability	Based on the content the data is ready to believe.	[11], [31]
Accuracy	The extent to which data seems reliable and error free.	[32], [33]
Objectivity	The degree to which the data is fair enough based on the given objectives.	[34]
Reputation	The status of the data should be high based on the generated source.	[11], [35]

Table II representing the second major dimension proposed in the framework known as "Contextual Data Quality". The main purpose of this factor is to assess the data based on its context. There are five different variables extracted to evaluate this data quality. All of the variables can be used to define and assess the background of dataset. It will help to the user to understand the data from its context. For example, the "timeliness" variable will provide the latest update of the data in the database. In this way, the data cannot be useful if it is showing very old statistics and employing on the current and active task. Altogether, there are 12 measuring items corresponding to the five variables to evaluate the contextual data quality.

TABLE II CONTEXTUAL DATA QUALITY - OPERATIONAL DEFINITION

Construct	Operational Definition	Source
Value-added	The extent to which the provided data is beneficial for the processing.	[36]
Relevancy	The degree to which the available data is relevant to the process.	[33]
Timeliness	The available data is updated and having time stamp.	[33]
Completeness	The data should be completed from all aspects and fulfil the requirements of the task.	[37]
Appropriate amount of data	The extent to which the available data should be appropriate based on volume and quantity.	[31], [38]

The data quality framework presented in this study has variety of attributes in order to measure the selected data in any decision making process. Table III represents the third data quality factor mentioned as "Representational Data Quality". This factor is used to assess the data based on its representation. This attribute is important as well to know the format and understanding level of the data set. There are four different variable extracted to assess this type of data quality, which supported by 8 measuring items. The quality metrics are defined in this research are based on the findings of literature review. Ease of understanding and interpretability are kind of variable selected in this study to evaluate the representational data quality. Ultimately, the results of this framework will support and improve the decision making process.

TABLE III REPRESENTATIONAL DATA QUALITY - OPERATIONAL DEFINITION

Construct	Operational Definition	Source
Interpretability	The degree of provided data should be in proper language with clear meta data.	[11], [36]
Ease of understanding	The extent to which data must be easy in understanding without any complexities.	[11], [36]
Representational consistency	The level of consistency between data transformation should be the same.	[11], [39]
Concise representation	The data should have proper formatting and representation.	[37]

Table IV elaborates the operational definition of each construct selected in this study for evaluating the "Accessibility Data Quality". The reason behind this quality measures is to understand the level of difficulty in getting the required data. There are two variables defined in this category known as "Accessibility" and "Access Security", which illustrate the both, security and accessibility level in attaining the data. The selected attributes has defined in the following table using operational definition, which highlight its use and demand in the framework.

TABLE IV ACCESSIBILITY DATA QUALITY - OPERATIONAL DEFINITION

Construct	Operational Definition	Source
Accessibility	The extent to which data is easily in access with proper requirements.	[25], [36]
Access Security	The level of security should be where data should be provided after proper validation.	[32], [33]

Finally, the last data quality that integrated in the framework is shown in Table V known as "Cognitive Data Quality". This data quality is used as intermediary between the main data quality variables and decision making process. The cognitive assessment is the major finding in this research to assess the data using human cognition theory. Overall, two variables used for cognitive assessment belong to user expertise on the given task and level of understanding in each step of the task. These two cognitive variables will help in the framework to understand the user's perspective on the data and given task.

TABLE V COGNITIVE DATA QUALITY – OPERATIONAL DEFINITION

Construct	Operational Definition	Source
User Expertise	The extent to which a user have expertise to understand data.	[26], [40]
Task Ambiguity	The level of understanding of a task assigned to the user.	[26], [41]

IV. CONCLUSION

Data quality assessment is very useful and essential for the organizations to improve their working capabilities, effective decision making process, and to run business processes with more liberty. The latest technologies resulted multiple applications and information systems that integrated well to perform organizational tasks. Therefore, the integration of the system required more concentration on maintaining the quality of the data from different perspectives. The centralized working environment and database need multiple directions to work on the data and to generate unique copy of the dataset. To deal in this situation, this research undertaken and important problem related with data quality assessment, which applied specifically on health industry. The research elaborates the step wise approach applied to achieve the main objective of the research. The main research methodology is based on proposing

framework. The proposed framework explained step by step to discuss all its major dimensions and measuring factors. Finally, the operational definition of each construct provided in detail, whereas the measuring items were also mentioned in this chapter. The proposed framework can be helpful for the organizations to enhance their data quality and decision making process. The validation of the framework can be applied in future to understand its applicability and implementation on real case study.

REFERENCES

[1] L. C. Günther, E. Colangelo, H. H. Wiendahl, and C. Bauer, "Data quality assessment for improved decision-making: A methodology for small and medium-sized enterprises," *Procedia Manuf.*, vol. 29, pp. 583–591, 2019.

[2] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, no. November, 2015.

[3] V. Vallurupalli and I. Bose, "Business intelligence for performance measurement: A case based analysis," *Decis. Support Syst.*, vol. 111, no. May, pp. 72–85, 2018.

[4] G. Marin, "Decision Support Systems," *Fac. Comput. Sci. Bus. Manag. , Rom. Am. Univ. , Bucharest , Rom.*, 2011.

[5] B. Heinrich, M. Klier, A. Schiller, and G. Wagner, "Assessing data quality – A probability-based metric for semantic consistency," *Decis. Support Syst.*, vol. 110, no. April, pp. 95–106, 2018.

[6] C. Batini and M. Scannapieco, "Data and information quality," *Cham, Switz. Springer Int. Publ. Google Sch.*, p. 43, 2016.

[7] Z. J. Gackowski, Q. U. D. Imensions, A. P. U. Ocused, and V. On, "Logical Interdependence of Some Attributes of Data/Information Quality.," in *ICIQ*, 2004, pp. 126–140.

[8] A. A. Abu-Musa, "Information technology and its implications for internal auditing," *Managerial Auditing Journal*, vol. 23, no. 5. pp. 438–466, 2008.

[9] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[10] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, "Overview and framework for data and information quality research," *J. Data Inf. Qual.*, vol. 1, no. 1, pp. 1–22, 2009.

[11] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manag. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.

[12] M. Ge and M. Helfert, "A review of information quality research—develop a research agenda," in *Paper presented at the International Conference on Information Quality 2007*, 2007.

[13] S.-T. Liaw *et al.*, "Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature," *Int. J. Med. Inform.*, vol. 82, no. 1, pp. 10–24, 2013.

[14] Y. Wang, "Decisions and Evaluation in Large Australian Companies Theory and Practice Compared," no. May, 2006.

[15] K.-U. Sattler, "THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT," *Encycl. Database Syst.*, pp. 1–5, 2016.

[16] H. Liu, W. Ke, K. Wei, and Z. Hua, "The impact of IT capabilities on firm performance: The mediating roles of absorptive capacity and supply chain agility," *Decis. Support Syst.*, vol. 54, no. 3, pp. 1452–1462, 2013.

[17] C. Fisher, E. Lauría, and S. Chengalur-Smith, *Introduction to information quality*. AuthorHouse, 2012.

[18] A. Haug, F. Zachariassen, and D. Van Liempd, "The costs of poor data quality," *J. Ind. Eng. Manag.*, vol. 4, no. 2, pp. 168–193, 2011.

[19] H.-T. Moges, K. Dejaeger, W. Lemahieu, and B. Baesens, "A multidimensional analysis of data quality for credit risk management: New insights and challenges," *Inf. Manag.*, vol. 50, no. 1, pp. 43–58, 2013.

[20] M. J. Carey *et al.*, "Data-Centric Systems and Applications," 2008.

[21] A. F. Karr, A. P. Sanil, and D. L. Banks, "Data quality: A statistical perspective," *Stat. Methodol.*, vol. 3, no. 2, pp. 137–173, 2006.

[22] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–52, 2009.

[23] L. Sebastian-Coleman, *Measuring data quality for ongoing improvement: a data quality assessment framework*. Newnes, 2012.

[24] H. Chen, D. Hailey, N. Wang, and P. Yu, "A review of data quality assessment methods for public health information systems," *Int. J. Environ. Res. Public Health*, vol. 11, no. 5, pp. 5170–5207, 2014.

[25] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: A methodology for information quality assessment," *Inf. Manag.*, vol. 40, no. 2, pp. 133–146, 2002.

[26] S. Watts, G. Shankaranarayanan, and A. Even, "Data quality assessment in context: A cognitive perspective," *Decis. Support Syst.*, vol. 48, no. 1, pp. 202–211, 2009.

[27] S. Raghunathan, "Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis," *Decis. Support Syst.*, vol. 26, no. 4, pp. 275–286, 1999.

[28] J. W. Creswell, *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Merrill, 2005.

[29] P. D. Leedy and J. E. Ormrod, *Practical Research: Planning and Design*, 10th ed. Pearson, 2013.

[30] U. Sekaran and R. Bougie, *Research Methods for Business: A Skill Building Approach*, 5th ed. John Wiley & Sons, 2010.

[31] F. Naumann and C. Rolker, "Assessment methods for information quality criteria," in *Proceedings of 5th International Conference on Information Quality*, 2000.

[32] R. H. J. Van Zeist and P. R. H. Hendriks, "Specifying software quality with the extended ISO model.," *Softw. Qual. J.*, vol. 5, no. 4, 1996.

[33] K. Kerr and R. Stockdale, "Data quality information and decision making: A healthcare case study," in *18th Australasian Conference on Information Systems*, 2007.

[34] J. E. Alexander and M. A. Tate, *Web wisdom: How to evaluate and create web page quality*. L. Erlbaum Associates Inc.., 1999.

[35] G. Shanks and B. Corbitt, "Understanding data quality: Social and cultural aspects," in *Proceedings of the 10th Australasian Conference on Information Systems*, 1999.

[36] A. Dedeke, "A conceptual framework for developing quality measures for information systems," in *Proceedings of 5th International Conference on Information Quality*, 2000.

[37] Z. Houhamdi and B. Athamena, "Impacts of information quality on decision-making," *Glob. Bus. Econ. Rev.*, vol. 21, no. 1, 2019.

[38] M. Helfert, O. Foley, M. Ge, and C. Cappiello, "Analysing the effect of security on information quality dimensions," in *ECIS*, 2009, pp. 2785–2797.

[39] P. Katerattanakul and K. Siau, "Measuring information quality of web sites: Development of an instrument.," in *Proceedings of the 20th international conference on Information Systems. Charlotte, North Carolina, United States*, 1999.

[40] K. Stamm and R. Dube, "The Relationship of Attitudinal Components to Trust in Media," *Communic. Res.*, vol. 21, no. 1, 1994.

[41] R. L. Daft and N. B. Macintosh, "A Tentative Exploration into the Amount and Equivocality of Information Processing in Organizational Work Units," *Adm. Sci. Q.*, vol. 26, 1981.