Empirical Validation of WebQMDW Model for Quality-based External Web Data Source Incorporation in a Data Warehouse

Priyanka Bhutani, Anju Saha, Anjana Gosain
USIC&T, GGSIP University, Dwarka, New Delhi, India

***Abstract*—In recent years, World Wide Web has emerged as the most promising external data source for organizations' Data Warehouses for valuable insights required in comprehensive decision making to gain a competitive edge. However, when the Data Warehouse uses external data sources from the Web without quality evaluation, it can adversely impact its quality. Quality models have been proposed in the research literature to evaluate and select Web Data sources for their integration in a Data Warehouse. However, these models are only conceptually proposed and not empirically validated. Therefore, in this paper, the authors present the empirical validation conducted on a set of 57 subjects to thoroughly validate the set of 22 quality factors and the initial structure of the multi-level, multi-dimensional WebQMDW quality model. The validated and restructured WebQMDW model thus obtained can significantly enhance the decision-making in the DW by selecting high-quality Web Data Sources.**

***Keywords*—Data warehouse; external data sources; web data sources; quality evaluation model; quality model validation**

I. Introduction

The importance of incorporating external data in the Data Warehouse to gain valuable insights into the market, competitors, products, or customers for comprehensive and unbiased decision making, has been long recognized in the research literature [1], [2]. The use of World Wide Web (WWW or Web) [3] as an external data source for the Data Warehouse (DW) [4] has grown considerably over the past few years [5]–[16]. The WWW helps provide a wide-angle lens for the decision-making in organizations in a very low-cost and highly accessible manner [4] (see Fig. 1). It is a known fact that the quality of the DW data sources hugely impacts the quality of the DW itself [1], [2]. This fact makes the quality-aware evaluation and selection of high quality, credible, and compatible Web Data Sources (WDSs) a very crucial task in the incorporation of Web data in the DW [4], [17]–[27]. There are, however, many challenges in this task like the availability of a massive amount of information, the heterogeneous structure and format of the Web Data [4], the dynamic nature [17], and poor reliability [18] of a significant chunk of Web Sources.

For the aforementioned task of quality-aware evaluation of Web Data Sources for a Data Warehouse, various quality models, frameworks, or a set of factors have been proposed in the research literature (see, for example, [19], [22]–[24],[4], [21], [20], [25], [26]). However, these quality models are only conceptual in nature. To the best of the authors' knowledge, none of these quality evaluation models for evaluating WDSs as external data sources for a DW have been empirically validated to corroborate their applicability in this problem area. In order to fill this research gap, in this paper, we present the empirical validation of the state-of-the-art multi-level, multi-dimensional WebQMDW (Web quality model for evaluating web sources for the DW) quality model [27] to enhance the decision making in a Data Warehouse. WebQMDW model [27] is the first of its kind model which segregates the quality factors in such a way to introduce automated quality evaluation as screening (at the first level) and separation of expert evaluation of different expert areas into different dimensions (at the second level). The present work complements and extends the authors' previous work [27] of the quality-based evaluation of the websites of academic institutions for incorporation as WDSs in a University DW. The said work proposed and used the novel $WSEM_{QT}$ (Web source evaluation with multi-criteria decision-making methods and web quality testing tools) process in conjunction with the underlying novel WebQMDW quality model. We believe that the empirical validation of the WebQMDW model will be an important milestone in the quality evaluation of WDSs for a DW, aiding the DW professionals in providing advanced data analytics for decision-making in the organization.

Hence, the objective of this paper is the empirical evaluation of the WebQMDW model in order to

- Validate the set of quality factors of the WebQMDW model; and eliminate or add new factors, if indicated by the validation results.
- Validate that the quality factors have been suitably placed in the level/dimension of the WebQMDW model; or if they should be placed in a different level/dimension according to the validation results.

The rest of the paper's overall arrangement is as follows: Section II discusses the frame of reference of the current work, including the related work, WebQMDW quality model, and motivation. Section III presents the empirical validation process of the WebQMDW model, including the analysis of results and restructuring of the model. Section IV discusses the various threats to the validity of the survey and how we dealt with them, followed by conclusion and future work in Section V.

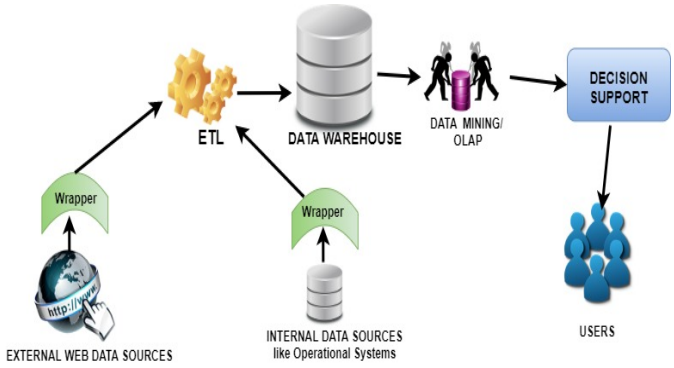


Fig. 1. Web Data as External Data Source for a Data Warehouse.

II. Frame of Reference

A. Related Work and the WebQMDW Quality Model

Quality is a critical and hence, widely researched concept in the context of both software products and data. From the point of view of Data/Information Quality (DQ/IQ), there are established standards (like ISO/IEC 25010/25012 [28], [29]) as well as "de facto" standards (like the Wang and Strong model [30]) in the relevant literature. Due to the peculiar characteristics of Web portals as opposed to a traditional software product, a plethora of research works have specifically addressed the Website/Web portal quality [31]–[33]. The quality evaluation of WDSs for a DW, however, encompasses the data quality as well as the source Website quality because due attention needs to be given to the quality requirements specific to the destination of the WDS incorporation, i.e., the Data Warehouse and the underlying business domain as well [27]. Few of the important works in the area of defining quality factors/models for using the WDSs as EDS for the DW are summarized in Table I. Huang et al. [4] have suggested integrating Web Data in a DW by considering both Quality and Coverage aspects. Quality aspect evaluation is proposed by using quality factors of Speed of loading, Accuracy, Currency, Presentation, Format, Content, and Source as put forward by Rieh [33]. Coverage aspect evaluation is proposed by determining two factors of Scope and Variety. Lóscio et al. [20] have used the three quality parameters of Data Completeness, Schema Completeness, and Correctness for determining the relevance of a WDS for a particular application domain. For a quality-aware Web Warehouse, Marotta et al. have proposed managing Data and Service Quality in their work [21]. In this work, the organization of Data quality is in six dimensions: Reliability, Consistency, Uniqueness, Freshness, Completeness, and Accuracy [21]; Whereas the organization of Service-Related quality is in six dimensions of Stability, Usability, Business Value, Security, Interoperability, and Service Level. The WebQM quality model proposed by Zhu and Buchman [19] has three classes of Web Source Stability, Web Application Specific Quality, and Web Information Quality, used to group the twelve quality factors in this model. These quality factors are Timeliness, Presentation, Relevance, Metadata, Objectivity, Completeness, Correctness, Origin, Refresh Rate, Durability, Accessibility, and Availability [19]. For WDS quality, Mihaila et al. have used the four quality factors: Granularity, Frequency of Updates, Recency, and Completeness. [25]. Naumann et al. used the three quality factors of Availability, Extent, and Understandability in their work [26].

In a previous work, authors have proposed the WebQMDW quality model [27] with 22 quality factors classified in 2 levels (Fig. 2). At Level-1(Automated quality testing level), those quality factors based on which the overall quality of the Website/Webpage can be assessed by using the available website quality testing automated tools are placed. This level has seven quality factors: Performance, Accessibility, Domain Reliability, SEO (Search Engine Optimization), Security, Best Practices, and Web Search Engine Ranking. At Level-2(Expert evaluation level), fifteen quality factors according to which the experts need to evaluate the WDSs are allocated. This level is further divided into three dimensions based on the expert area required for judging them. Dimension 1, with Web Data Related Quality Factors, has five quality factors: Interoperability, Media Format, Cost of Access, Amount of Data, and Timeliness. A Web Data expert evaluates them. Dimension 2, with Data Warehouse Context-related Quality Factors, has five quality factors: Metadata interpretability, Time Period Correspondence, Concise Representation, Consistent Representation, and Completeness. A DW expert evaluates them. Dimension 3, with Business Domain related Quality Factors, has five quality factors: Business Value Addition, Accuracy, Objectivity, Believability, and Uniqueness. A Business Domain Expert evaluates them. In the current work, we choose to focus on this model as the selection and structuring of quality factors is in such a way that it solves the two main issues of the quality evaluation process of Web data sources [27]. The first issue of an enormous load of evaluation on experts is tackled due to the screening of the vast number of web sources to a select few through automated Website quality testing tools at the 1st level of the model. The second issue in previous quality models was the bottleneck of finding experts with expertise in all the related domains of quality evaluation. This issue is also resolved at the 2nd level of the model as different experts need to evaluate the quality factors separated into different dimensions according to the required expertise, namely Web, Data Warehouse, and underlying business domain. The initial structure of the WebQMDW model is shown in Fig. 2. The details of these 22 quality factors in the context of the Web Source Quality evaluation and the detailed account of the model's application to a practical case study of a University DW can be found in [27].

TABLE I. SUMMARY OF SOME WORKS DEFINING QUALITY MODELS FOR QUALITY-BASED EVALUATION OF WDSS FOR DW

Author(s)	Domain	Structure of quality model/ framework	Whether empirically validated? (Yes/No)
Huang et al. [4]	Quality of WDS for DW	2 dimensions; Total 9 QFs	No
Lóscio et al. [20]	Web Data Source quality	Total 3 QFs	No
Marotta et al. [21]	Quality-aware Web Warehouse	2 categories; Total 12 QFs	No
Zhu and Buchman [19]	Quality of WDS for DW	3 categories; Total 12 QFs	No
Mihaila et al. [25]	WWW Source selection	Total 4 QFs	No
Naumann et al. [26]	Quality driven Web Source selection	3 dimensions; Total 3 QFs	No
Bhutani et.al [27]	Quality of WDS for DW	2 levels, 3 dimensions; Total 22 QFs	No

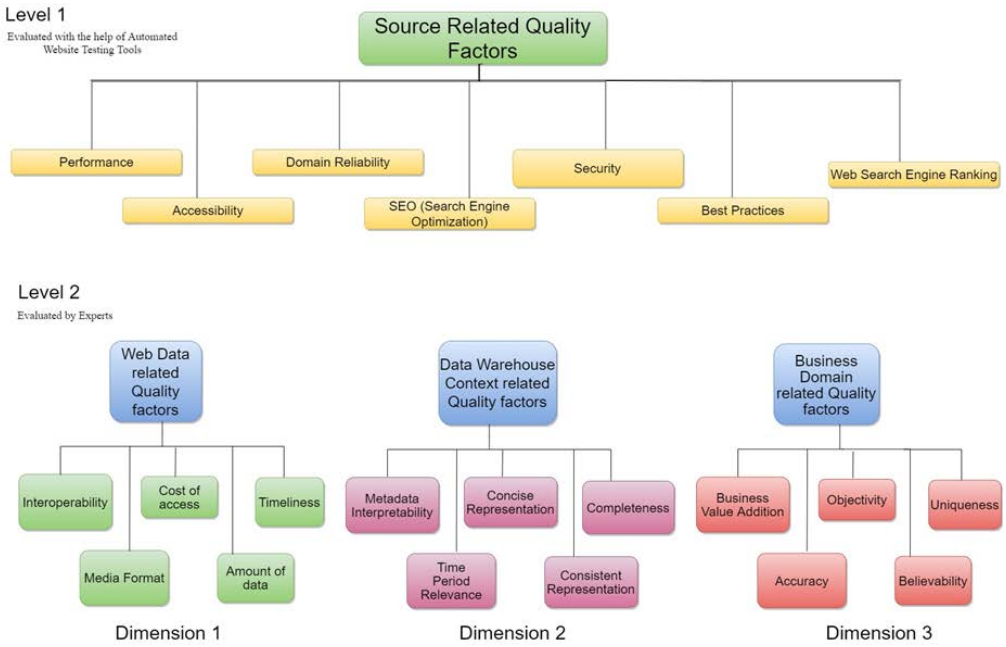


Fig. 2. WebQMDW- a Quality Evaluation Model for Web Data Sources as EDS of a DW [36].

B. Motivation

For any area, after proposing the quality model, the next step is its validation from the perspective of the users of the respective domain. The validation is important due to two reasons. Firstly, it is essential to consider users' choices of quality factors due to their profile, experience, and knowledge in the respective field [34]. Secondly, the robust statistical analysis from the empirical validation attaches confidence to the adequacy of the proposed quality model [35].

In the area of generic quality evaluation of Websites, there are several empirical validations works in the research literature (Table II). The work of Caro et al. [32] validates a quality model for evaluation of the quality of Websites. The research work [36] of Moraga et al. validates a quality model for website quality specifically from the point of view of "University-educated users." A quality model for evaluating health websites' data quality is validated through experts in the work of Liete et al. [35]. Some authors [32], [36] have performed the validation using the survey method. Few authors have used the Delphi method [37] for the validation of the quality models [35].

However, no such validation works of quality models specific to the context of evaluation of WDSs as EDS to a DW (see Table I) could be found in the research literature, to the best of the authors' knowledge and belief. Hence, the current work provides the validation of the WebQMDW model (described in the previous section), which is specific for the said context [27]. It is performed by using the survey method according to the guidelines of the work of Pfleeger and Kitchenham[38]–[43], as described in the subsequent sections (see Sections III and IV).

TABLE II. SUMMARY OF SOME RELATED WORKS DEFINING/ VALIDATING QUALITY MODELS FOR EVALUATION OF WEBSITES

Author(s)	Domain	Structure of quality model/ framework	Whether Validated? (Yes/No)
Caro et al. [32]	Web Portal Quality	4 categories; Total 33 QFs	Yes
Moraga et al. [36]	Web Portal Quality	4 categories; Total 42 QFs	Yes
Liete et al. [35]	Health UnitWebsites' Quality	3 categories; Total 23 QFs	Yes

III. Validation Process of the WebQMDW Model

As described in the previous section, the WebQMDW model has been obtained by bearing in mind the definitions of the quality factors and the defined categories (i.e., levels/dimensions) identified for structuring the factors. This section elaborates the validation process of both the set of quality factors and the initial hierarchical structure of the WebQMDW model.

A. Research Methodology

Several empirical validation methods [44] are described in the research literature, like case studies, controlled experiments, ethnographies, and surveys. The survey method [38]–[44] has a defining characteristic of studying the applicability of the phenomenon on the target population by polling the survey questionnaire on the representative subset of the target population. Bearing in mind that we need to study the applicability of the WebQMDW model in the opinion of the Web and Data Warehouse users (the target population), this was the best applicable method for our work. So, in this paper, we use the survey method as the validation method to thoroughly validate the set of quality factors and the structure of the WebQMDW model while following the guidelines and principles of research proposed by Kitchenham and Pfleeger [38]–[43]. These guidelines describe the various activities for collecting information for describing, comparing, or explaining knowledge, behavior, and attitudes, by using the survey instrument [43].

B. Setting of Objectives

Measurable and specific objectives are set in this step. We set the main objective of our survey as: ‘‘To acquire the viewpoint of Web and Data Warehouse users regarding the importance as well as the placement (in the levels/dimensions) of each of the quality factors in the WebQMDW model.”

C. Selection of Subjects

Keeping the objective mentioned earlier in mind, the target subjects required were both Web and Data Warehouse users. For the purpose of empirical analysis, many researchers have pointed out the advantages of taking students as subjects [45], [46] as the students’ knowledge tends to be homogenous and a high number of students as subjects are conveniently available simultaneously. According to us, the students who have the knowledge and practical hands-on experience of Data Warehouse and the World Wide Web were well suited to be this survey’s subjects. Additionally, if the students have the knowledge of Data and Software Quality, they will be able to assess the importance of each quality factor better. Hence, it was decided to use “Convenience sampling” and administer the survey to a set of students, the 74 students of the Data Warehousing & Mining Course of the final-year class of Information Technology program at USIC&T, GGSIP University, New Delhi, India. These students not only had knowledge of the Web and Data Warehouse but had also studied an entire course on Software Engineering as part of their curriculum previously. The survey was conducted as a part of the mandatory practical laboratory session of the Data Warehousing & Mining course. Therefore, there was enough motivation in the students to be a part of the survey.

D. Selection of the Design of the Survey

The descriptive design of the survey is considered most appropriate, where the objective requires a description of the phenomenon of interest. The objective of this survey requires a description of the opinion of the respondents regarding the importance and placement of quality factors in the WebQMDW model. Hence, the descriptive design [38] was considered appropriate and selected by us rather than the experimental design.

E. Preparation of the Survey Instrument

The guidelines of designing the survey instrument, i.e., the questionnaire [39], suggest that the survey questions should be chosen, keeping in mind the objective of the survey. Hence, in accordance with the objective mentioned earlier, we constructed the questionnaire with 22 Likert-style closed questions divided into sections I and II, asking the importance of the 22 quality factors of WebQMDW model Level I and Level II, respectively (Fig. 3, Fig. 4). Only the naming of quality factors in the questions could have led to ambiguity in the respondents' minds about the meaning of the quality factors. So, we formulated the questions in conventional simple English language by adapting the definition of each factor from the research literature [27], [32]. The answers to the closed questions were supposed to be marked in the 5-point Likert scale ranging from the lowest score ‘1’ signifying ‘Not Important’ and highest score ‘5’ signifying ‘Very Important.’ Section III consisted of 2 open questions regarding the structural placement of quality factors in the levels/dimensions of the WebQMDW model (Fig. 5). The first open question focused on any suggested switching of the category (i.e., Level/Dimension) of the factors in the WebQMDW model. The second open question focused on any other quality aspect or factor to be added to the WebQMDW model.

Section- I (corresponding to importance of WebQMDW Level 1 quality factors)

Level 1

Q.1 The importance value of the Performance i.e the Speed of loading of the Web Source, should be:
Q.2 The importance value of the Web Source having proper navigation mechanisms to be accessed speedily and with ease, should be:
Q.3 The importance value of the Web Source domain being considered trustworthy and delivering appropriate data, should be:
Q.4 The importance value of the Web Source having a strong SEO (Search Engine Optimization) for the relevant data to be discovered easily, should be:
Q.5 The importance value of the Web Source having security provisions (like SSL certificate) for preventing manipulation and unauthorized access to data, should be:
Q.6 The importance value of the various Best Practices that are followed by the Web Source (e.g practice of deferring download of unnecessary resources), should be:
Q.7 The importance value of the Web Source having high popularity and being considered worthy of great reputation for its content and services, should be:

Fig. 3. Survey Questionnaire -Section I (Reproduction of Questions from the Google form Questionnaire).

Section- II (corresponding to importance of WebQMDW Level 2 quality factors)

Level 2, Dimension 1
Q.8 The importance value of the Web Source data having the ability to be accessible over different platforms (operating systems or hardware architecture), should be:
Q.9 The importance value of the media format (text/HTML/pdf/audio/video etc) of the data from the Web Source fitting within the processing ability of the organization, should be:
Q.10 The importance value of the degree to which the data from the Web source is worthy of the cost associated (if it requires access fees), should be:
Q.11 The importance value of the amount or quantity of data provided by the Web Source being significant, should be:
Q.12 The importance value of the Web Source providing the data within the time constraint specified by the need of organization, should be:

Level 2, Dimension 2:
Q.13 The importance value of the description or metadata of the data from the Web Source being easy to interpret in accordance with the Data Warehouse schema, should be:
Q.14 The importance value of the data from the Web Source corresponding to the required time period according to the usage of Web data in Data Warehouse, should be:
Q.15 The importance value of the data from the Web Source being concise and free of superfluous elements that are not required for the right purpose in the Data Warehouse, should be:
Q.16 The importance value of the data from the Web Source being consistently represented in same or compatible formats throughout the Web pages of the Web Source, should be:
Q.17 The importance value of the data of the Web Source providing a complete coverage in terms of the depth, breadth and scope of the task at hand of the Data Warehouse, should be:

Level 2, Dimension 3:
Q.18 The importance value of the degree to which the data from Web Source is beneficial and adds value to the business of the organization, should be:
Q.19 The importance value of the data from the Web Source being correct and guaranteed to be error-free especially in the context of the application domain, should be:
Q.20 The importance value of the data from the Web Source being impartial and free from bias, should be:
Q.21 The importance value of the extent to which the data from the Web Source is believable, should be:

Fig. 4. Survey Questionnaire -Section II (Reproduction of Questions from the Google form Questionnaire).

Section- III (corresponding to open questions for structure of WebQMDW model)

Q.23 Do you suggest the switching of the category (i.e level or dimension in the WebQMDW quality model) of any quality factor?
Q.24 Do you suggest the addition of any new quality factor, not covered in the WebQMDW quality model?

Fig. 5. Survey Questionnaire – Section III (Reproduction of Questions from the Google form Questionnaire).

F. Validation of the Survey Instrument

We pre-tested the questionnaire to validate the survey instrument. Ten respondents (5 of them pursuing Ph.D. in the field of Data Warehousing and the rest 5 pursuing Ph.D. in the field of Web Engineering) answered the questionnaire before its actual administration. Following their feedback about the understanding of the questions, three questions (with questions no. 6, 9, and 10) were modified with examples and simpler language to improve the questionnaire.

G. Administration of Survey

The survey was administered to the subjects in an online session of a Data Warehousing & Mining laboratory class. The questionnaire was delivered in the form of a Google form whose link was shared with the subjects. Before the beginning of the session, the purpose and importance of the study were briefly explained to the respondents. The time limit of one hour for submitting the responses to the survey was also communicated to them.

H. Analysis of the Data

The survey was supposed to be administered to an expected sample of 74 subjects. In the actual scenario, the survey session was attended by 59 subjects because the remaining subjects were absent during the session. The recorded response rate was, hence, 79.7%. However, during the session, two subjects could not complete the survey due to network issues. So, the rest of the 57 responses were considered.

First, we analyzed the internal consistency of our data from closed questions with the help of Cronbach's alpha value (Equation 1) [47]. We determined Cronbach's alpha for data of section I and section II of the questionnaire, corresponding to importance value responses for quality factors from Level 1 and Level 2 of the WebQMDW model, respectively (see Table III). As a thumb rule, the value of Cronbach's alpha above 0.7 is considered acceptable. For our data, the values obtained were 0.889 for Section I and 0.920 for Section II. Hence, the survey can be said to have good internal consistency and reliable results for further analysis.

TABLE III. RESULTS OF CRONBACH'S ALPHA VALUE ANALYSIS

Questionnaire Data	**Cronbach's Alpha Value**
Section I (corresponding to Level 1 of WebQMDW model)	0.889
Section II (corresponding to Level 2 of WebQMDW model)	0.920

Cronbach's alpha [47] i.e., $\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$ (1)

Where N= the number of items

$\bar{c}$= average inter-item covariance

$\bar{v}$= average variance

Tables IV and V shows the descriptive statistics of the responses for the importance values for the 22 quality factors. In this work, we have calculated the mean (i.e., average value) and the percentage coefficient of variation (%CV) of the importance values, to be used as the indicators for including or excluding the quality factors. It was decided to eliminate the factors whose mean value was below the value 3.0 (mid-point of the scale) as conceptually, in the view of the participants, they did not seem significant enough to be considered a quality factor for the evaluation of Web sources. We also decided to eliminate those quality factors for whom the percent variation coefficient was above 33% because conceptually, there was inconsistency in the participants' viewpoint about the importance of this quality factor. Thus, considering the ranked values of mean importance of factors in Fig. 6, most of the 22

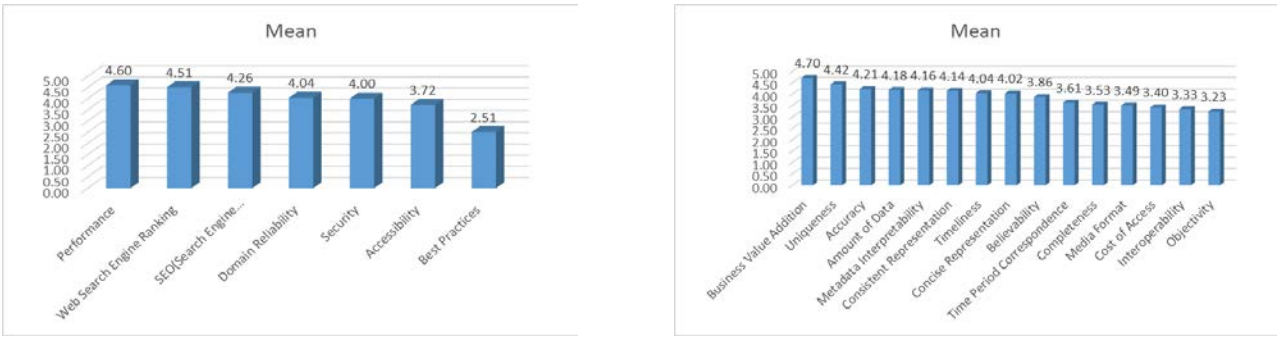
factors of the WebQMDW model had a mean importance value above 3. These values signified that the respondents considered most of the factors to be having moderate or high importance. Among the highly important factors were Performance, Web Search Engine Ranking, Business Value Addition, and Uniqueness. However, the quality factor Best Practices was eliminated as its mean value fell below the decided indicator of 3.0. None of the factors had a percent variation coefficient of above 33%, so no factor was eliminated for this particular constraint (Fig. 7). The open question (number 23), which focused on any suggested switching of the category (i.e., level or dimension) of any quality factor, was not answered by any of the respondents. The last open question (number 24), which focused on the addition of any new quality factor, was answered by four participants who suggested including Reputation as one of the factors. On close review of meanings of the factors from the review of literature, it was seen that in the context of Web Sources, in particular, this factor of Reputation [30] was synonymous to the factor Web Search Engine Ranking that was already included in the WebQMDW model [27]. The factor name Reputation was also used in the pioneering work of Wang and Strong, considered a de-facto Data Quality standard [30]. So, instead of adding another factor, we decided to consider renaming the factor Web Search Engine Ranking to the more general name of Reputation.

TABLE IV. DESCRIPTIVE STATISTICAL ANALYSIS OF EACH QUALITY FACTOR OF LEVEL 1

Quality Factor	Min. Value	Max. Value	Mean Value	Standard Deviation	%CV
Performance	3	5	4.60	0.53	14.50%
Accessibility	2	5	3.72	0.80	21.41%
Domain Reliability	3	5	4.04	0.57	14.02%
SEO (Search Engine Optimization)	2	5	4.60	0.77	18.02%
Security	3	5	4.00	0.46	11.57%
Best Practices	1	4	2.51	0.71	28.31%
Web Search Engine Ranking	2	5	4.51	0.68	15.18%

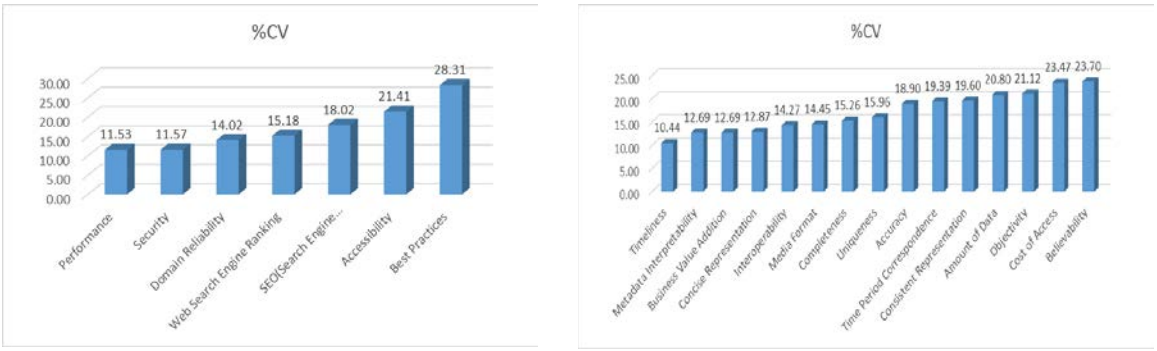
TABLE V. DESCRIPTIVE STATISTICAL ANALYSIS OF EACH QUALITY FACTOR OF LEVEL 2

Dimension	Quality Factor	Min. Value	Max. Value	Mean Value	Standard Deviation	%CV
Dimension 1	Interoperability	3	4	3.33	0.48	14.27
	Media Format	3	4	3.49	0.50	14.45
	Cost of Access	2	5	3.40	0.80	23.47
	Amount of Data	2	5	4.18	0.87	20.80
	Timeliness	3	5	4.04	0.42	10.44
Dimension 2	Metadata Interpretability	3	5	4.16	0.53	12.69
	Time Period Correspondence	2	5	3.61	0.70	19.39
	Concise Representation	3	5	4.02	0.52	12.87
	Consistent Representation	2	5	4.14	0.81	19.60
	Completeness	3	5	3.53	0.54	15.26
Dimension 3	Business Value Addition	2	5	4.70	0.60	12.69
	Accuracy	3	5	4.21	0.80	18.90
	Objectivity	1	4	3.23	0.68	21.12
	Believability	2	5	3.86	0.91	23.70
	Uniqueness	2	5	4.42	0.71	15.96



(a) (b)

Fig. 6. Ranked Quality Factors of WebQMDW Model According to Mean of Importance Values (a) Level-1 (b) Level-2.



(a) (b)

Fig. 7. Ranked Quality Factors of WebQMDW Model According to Percent Coefficient of Variation (%CV) of Importance Values (a) Level-1 (b) Level-2.

I. Restructuring of the Quality Factors of the WebQMDW Model

The initial structure of the WebQMDW model is shown in Fig. 1. After completing the above-stated validation process, the WebQMDW model now consists of a set of 21 factors, instead of 22, as one of the factors Best Practices was eliminated in the validation. As stated above, the factor Web. Search Engine Ranking was renamed as Reputation. Since none of the participants suggested switching of the categories (i.e., level/dimension) of the factors, no other restructuring was done. The final structure of the WebQMDW model (with the above-stated changes) is as shown in Fig. 8.

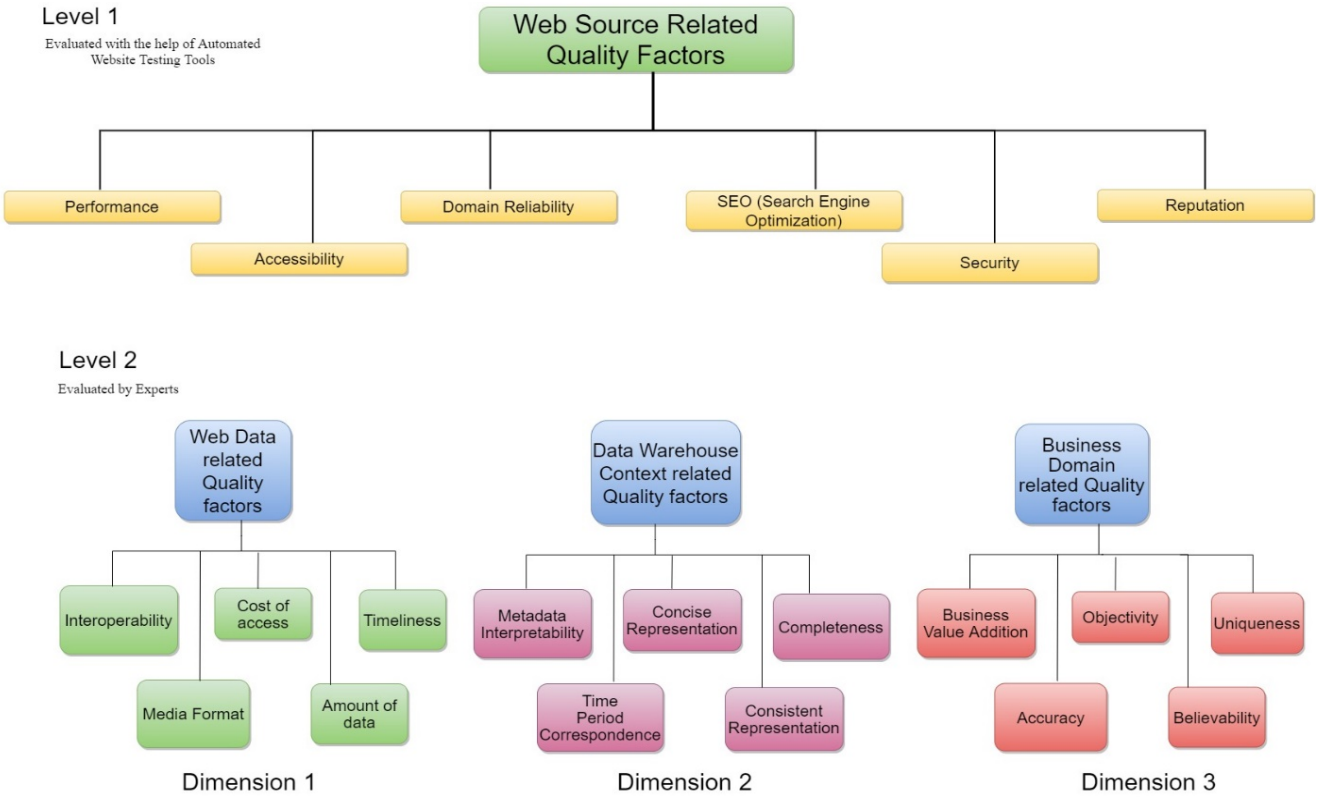


Fig. 8. Restructured WebQMDW Model- after the Validation Process.

IV. Threats to Validity

This section discusses the threats to the following kinds of validity and also how they were minimized:

A. Construct Validity

The survey uses the 5-point Likert scale to gather the opinion of the participants about the importance of the factors, with the lowest numerical value '1' signifying 'Not Important' and the highest value '5' signifying 'Very Important.' The Likert scale is used in many previous similar studies [32], [35] to gather the opinion of participants. This scale is an efficient tool for observation and hence, can be considered as a valid construct.

B. Internal Validity

To ensure internal validity is to make sure that the results are not being derived from casual relationships. For this aspect, we considered the following issues carefully:

- The students enrolled in the same class of Data Warehousing & Mining were taken as subjects. The subjects had adequate knowledge of Data and Software Quality as they had also studied an entire course on Software Engineering as part of their curriculum. Hence, it can be said that all the subjects had the same profile and level of experience both in Data Warehousing and Data Quality. Thus, the variability among subjects was reduced.
- Since the subjects had not taken part in any survey on the same lines as the present one, so the persistence effect was nullified.
- Since the survey was provided to be filled only once, so no learning could have taken place. Thus, the threat of the learning effect was not present.
- The survey was administered in a one-hour session. This time was much less than even one practical laboratory session time of the students. Hence, the fatigue effect was not that relevant in this case.
- The survey was conducted as part and parcel of the ongoing practical laboratory sessions of the subjects' Data Warehousing & Mining course. The subjects were also motivated by telling them the importance of their contribution to the current research in the Data Warehousing field. Also, since subjects had already studied Web Warehousing as one of the advanced topics in the course, they showed sufficient interest in participating. Hence, we had achieved sufficient subject motivation for the survey.
- Since the survey was conducted in an online session with the subjects participating from their homes, their influence on each other was, if at all, very minimal. Further, to avoid plagiarism, it was ensured that the subjects kept their videos on during the entire one-hour session and were informed not to communicate with each other.

C. External Validity

External validity is the degree of generalizability of the research results to the population of interest and beyond in actual practice. External validity was ensured by mitigating the following two issues:

- Material and task used: A survey questionnaire structured as a Google form was the material used. This survey was independent as no previous task was needed to be done in order to fill it.
- Subjects: The students were used as subjects of this survey due to two major reasons. Firstly, the students clearly represented the population understudy for the survey as they had experience as Data Warehouse users as well as Web Portal users, along with the knowledge of Data Quality. Secondly, many researchers have argued in favor of using students as subjects [45], [46] without impacting the external validity much. However, we do not rule out the possibility of conducting a replicated study with experts from the industry in the near future.

D. Conclusion Validity

Conclusion validity is the statistical validity of the conclusion of the research. For this concern, the size of the sample (57 subjects) could be the only issue. However, most of the quality factors identified from the research literature have been previously used and mostly validated, in the sub-areas of the current problem domain, like Web Portal quality and Data Warehouse quality. Hence, the concern is subjugated. We will still consider conducting a replication study with a larger number of subjects from the industry.

V. Conclusion and Future Work

Over the last few decades, Web Data Sources have established their position as good, viable, and highly accessible External Data Sources for a Data Warehouse. However, the assessment of the quality of the Web Data Source is critical before their incorporation in the DW. Some quality models have been conceptually proposed in the research literature. However, to the best of the authors' knowledge, none of the previously known models for the Web Data Source evaluation for a Data Warehouse have been empirically validated. Hence, this paper presents the validation process of the multi-level, multi-dimensional WebQMDW model for quality evaluation of Web data sources for a Data Warehouse. The objective was to provide an empirically validated quality model which will guide the DW professionals to provide enhanced decision making in the Data Warehouse by quality-based incorporation of external Web data sources. The thorough empirical validation is carried out through a survey based on the Pfleeger and Kitchenham work guidelines, which are considered a de-facto standard. A questionnaire with three sections was used as the instrument for the survey. Sections I and II correspond to the importance values of the quality factors from level 1(automated quality evaluation) and level 2(expert evaluation) of the WebQMDW model. Section III focuses on the structuring of the model into levels and dimensions. The statistical analysis of the results obtained from the validation survey revealed that 21 factors of the WebQMDW model are

considered to be having either high or moderate importance for Web Source quality evaluation. The restructured and validated WebQMDW was obtained as suggested by the results of the empirical validation and supported by the research literature, which can be considered a significant contribution in this area. We plan to conduct a further study with a larger number of subjects, especially from the industry, in the near future. Such a study could be beneficial to refine the model further. We also plan to work on the measures for each quality factor and the refining of the granularity of the model.

REFERENCES

[1] M. Strand, "External Data Incorporation into Data Warehouses, Doctoral Thesis,Department of Computer and System Sciences, Stockholm University," Department of Computer and System Sciences, Stockholm University, Oct. 2005.

[2] M. Niklasson, "Problems Concerning External Data Incorporation in Data Warehouses,Dissertation for the degree of M.Sc.,The school of Humanities and Informatics, University of Skovde," The school of Humanities and Informatics, University of Skovde, 2004.

[3] P. Bhutani and A. Saha, "Towards an Evolved Information Food Chain of World Wide Web and Taxonomy of Semantic Web Mining," in International Conference on Innovative Computing and Communications, vol. 56, S. Bhattacharyya, A. E. Hassanien, D. Gupta, A. Khanna, and I. Pan, Eds. Singapore: Springer Singapore, 2019, pp. 443–451. doi: 10.1007/978-981-13-2354-6_46.

[4] Z. Huang, L.-D. Chen, and M. N. Frolick, "Integrating Web-Based Data into A Data Warehouse," Information Systems Management, vol. 19, no. 1, pp. 23–34, Jan. 2002, doi: 10.1201/1078/43199.19.1.20020101/31473.4.

[5] A. Alrefae and J. Cao, "Intensional XML-enabled web-based real-time decision support system," in 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, Oct. 2017, pp. 1–10. doi: 10.1109/ICCNI.2017.8123819.

[6] F. Ravat and J. Song, "Enabling OLAP analyses on the web of data," in 2016 Eleventh International Conference on Digital Information Management (ICDIM), Porto, Portugal, Sep. 2016, pp. 215–224. doi: 10.1109/ICDIM.2016.7829762.

[7] R. V. Nikam, S. Shirwaikar, and V. S. Kharat, "Conceptual model for a data warehouse on the web," in 2016 IEEE Bombay Section Symposium (IBSS), Baramati, India, Dec. 2016, pp. 1–6. doi: 10.1109/IBSS.2016.7940201.

[8] R. Mehmood, M. U. Shaikh, R. Bie, H. Dawood, and H. Dawood, "IoT-enabled Web warehouse architecture: a secure approach," Pers Ubiquit Comput, vol. 19, no. 7, pp. 1157–1167, Oct. 2015, doi: 10.1007/s00779-015-0882-8.

[9] Y. Jiang, Z. Shao, Y. Guo, H. Zhang, and L. Sun, "Building XML Data Warehouse with Data Reconstruction by Knowledge Graph," in 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, Dalian, China, 2015, pp. 314–320. doi: 10.1109/BDCloud.2015.48.

[10] A. Delgado and A. Marotta, "Automating the process of building flexible Web Warehouses with BPM Systems," in 2015 Latin American Computing Conference (CLEI), Arequipa, Peru, Oct. 2015, pp. 1–11. doi: 10.1109/CLEI.2015.7360005.

[11] R. Mehmood, M. U. Shaikh, L. Ma, and R. Bie, "Enhanced Web Warehouse Model: A Secure Approach," in 2014 International Conference on Identification, Information and Knowledge in the Internet of Things, Beijing, China, Oct. 2014, pp. 88–91. doi: 10.1109/IIKI.2014.26.

[12] S. Destercke, P. Buche, and B. Charnomordic, "Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse," IEEE Trans. Knowl. Data Eng., vol. 25, no. 1, pp. 92–105, Jan. 2013, doi: 10.1109/TKDE.2011.179.

[13] L. G. Moya, S. Kudama, M. J. A. Cabo, and R. B. Llavori, "Integrating web feed opinions into a corporate data warehouse," in Proceedings of the 2nd International Workshop on Business intelligencE and the WEB - BEWEB '11, Uppsala, Sweden, 2011, pp. 20–27. doi: 10.1145/1966883.1966891.

[14] O. Boussaid, J. Darmont, F. Bentayeb, and S. Loudcher, "Warehousing complex data from the web," International Journal of Web Engineering and Technology, vol. 4, no. 4, pp. 408–433, Jan. 2008, doi: 10.1504/IJWET.2008.019942.

[15] L. Yu, W. Huang, S. Wang, and K. K. Lai, "Web warehouse – a new web information fusion tool for web mining," Information Fusion, vol. 9, no. 4, pp. 501–511, Oct. 2008, doi: 10.1016/j.inffus.2006.10.007.

[16] A. Marotta, R. Motz, and R. Ruggia, "Managing source schema evolution in web warehouses," J. Braz. Comp. Soc., vol. 8, no. 2, pp. 20–31, Nov. 2002, doi: 10.1590/S0104-65002002000200003.

[17] E. A. Rundensteiner, A. Koeller, and X. Zhang, "Maintaining data warehouses over changing information sources," Commun. ACM, vol. 43, no. 6, pp. 57–62, Jun. 2000, doi: 10.1145/336460.336475.

[18] H. Keshavarz, "How Credible is Information on the Web: Reflections on Misinformation and Disinformation," Infopreneurship Journal, vol. 1, no. 2, pp. 1–17, 2014.

[19] Yan Zhu and A. Buchmann, "Evaluating and selecting Web sources as external information resources of a data warehouse," in Proceedings of the Third International Conference on Web Information Systems Engineering, 2002. WISE 2002., Singapore, 2002, pp. 149–160. doi: 10.1109/WISE.2002.1181652.

[20] B. F. Lóscio, M. C. Batista, D. Souza, and A. C. Salgado, "Using information quality for the identification of relevant web data sources: a proposal," in Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, 2012, pp. 36–44.

[21] A. Marotta, L. González, and R. Ruggia, "A quality aware service-oriented web warehouse platform," in Proceedings of the 2012 Joint EDBT/ICDT Workshops on - EDBT-ICDT '12, Berlin, Germany, 2012, p. 29. doi: 10.1145/2320765.2320783.

[22] H. S. Sinha, "Enhancement of TOPSIS for Evaluating the Web-Sources to Select as External Source for Web-Warehousing," IJRSDA, vol. 5, no. 1, pp. 117–130, Jan. 2018, doi: 10.4018/IJRSDA.2018010108.

[23] Y. Zhu, "Group Assessment of Web Source/Information Quality Based on WebQM and Fuzzy Logic," in Rough Sets and Knowledge Technology, Berlin, Heidelberg, 2008, pp. 660–667.

[24] Y. Ding, Q. Li, and Y. Dong, "Web Source Evaluation and Selection by Mass Collaboration," in 2009 Second International Workshop on Knowledge Discovery and Data Mining, Moscow, Russia, Jan. 2009, pp. 741–744. doi: 10.1109/WKDD.2009.71.

[25] G. A. Mihaila, L. Raschid, and M. E. Vidal, "Using quality of data metadata for source selection and ranking," in Proceedings of the Third International Workshop on the Web and Databases, WebDB, May 2000, pp. 93–98.

[26] F. Naumann, J. Freytag, and M. Spiliopoulou, "Quality Driven Source Selection Using Data Envelope Analysis.," USA, 1998, p. 152.

[27] P. Bhutani, A. Saha, and A. Gosain, "WSEM QT : a novel approach for quality-based evaluation of web data sources for a data warehouse," IET softw., vol. 14, no. 7, pp. 806–815, Dec. 2020, doi: 10.1049/iet-sen.2020.0088.

[28] "ISO/IEC 25010:2011," Systems and software Quality Requirements and Evaluation (SQuaRE). System and software quality models, 2011. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/57/35733.html (accessed Jun. 19, 2021).

[29] "ISO/IEC 25012:2008," Software product Quality Requirements and Evaluation(SQuaRE) - Data quality model, 2008. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/57/35736.html (accessed Jun. 19, 2021).

[30] R. Y. Wang, "A product perspective on total data quality management," Commun. ACM, vol. 41, no. 2, pp. 58–65, Feb. 1998, doi: 10.1145/269012.269022.

[31] A. Caro, C. Calero, I. Caballero, and M. Piattini, "A First Approach to a Data Quality Model for Web Portals," in Computational Science and Its Applications - ICCSA 2006, Berlin, Heidelberg, 2006, pp. 984–993.

[32] A. Caro, C. Calero, I. Caballero, and M. Piattini, "A proposal for a set of attributes relevant for Web portal data quality," Software Qual J, vol. 16, no. 4, pp. 513–542, Dec. 2008, doi: 10.1007/s11219-008-9046-7.

[33] S. Y. Rieh, “Judgment of information quality and cognitive authority in the Web,” Journal of the American Society for Information Science and Technology, vol. 53, no. 2, pp. 145–161, Jan. 2002, doi: 10.1002/asi.10017.

[34] J. Ruževi, “Peculiarities of the Business Information Quality Assessment,” Vadyba [Vilniaus universitetas], vol. Nr. 1, no. 14, pp. 54-60., 2007.

[35] P. Leite, J. Gonçalves, P. Teixeira, and Á. Rocha, “A model for the evaluation of data quality in health unit websites,” Health Informatics J, vol. 22, no. 3, pp. 479–495, Sep. 2016, doi: 10.1177/1460458214567003.

[36] C. Moraga, M. Á. Moraga, A. Caro, R. R. Muñoz, and C. Calero, “University educated users’ data quality preferences in web portals,” IJIQ, vol. 3, no. 2, pp. 107–126, 2013, doi: 10.1504/IJIQ.2013.054274.

[37] F. Hasson, S. Keeney, and H. McKenna, “Research guidelines for the Delphi survey technique: Delphi survey technique,” Journal of Advanced Nursing, vol. 32, no. 4, pp. 1008–1015, Oct. 2000, doi: 10.1046/j.1365-2648.2000.t01-1-01567.x.

[38] B. Kitchenham and S. L. Pfleeger, “Principles of Survey Research Part 2: Designing a Survey,” ACM SIGSOFT Software Engineering Notes, vol. 27, no. 1, pp. 18–20, 2002.

[39] B. Kitchenham and S. L. Pfleeger, “Principles of survey research: part 3: constructing a survey instrument,” SIGSOFT Softw. Eng. Notes, vol. 27, no. 2, pp. 20–24, 2002, doi: 10.1145/511152.511155.

[40] B. Kitchenham and S. L. Pfleeger, “Principles of Survey Research Part 4: Questionnaire Evaluation,” ACM SIGSOFT Software Engineering Notes, vol. 27, no. 3, pp. 20–23, 2002.

[41] B. Kitchenham and S. L. Pfleeger, “Principles of survey research: part 5: populations and samples,” SIGSOFT Softw. Eng. Notes, vol. 27, no. 5, pp. 17–20, 2002, doi: 10.1145/571681.571686.

[42] B. Kitchenham and S. L. Pfleeger, “Principles of Survey Research Part 6: Data Analysis,” ACM SIGSOFT Software Engineering Notes, vol. 28, no. 2, pp. 24–27, 2003.

[43] B. Kitchenham and S. L. Pfleeger, “Principles of survey research: part 1: turning lemons into lemonade,” SIGSOFT Softw. Eng. Notes, vol. 26, no. 6, pp. 16–18, Nov. 2001, doi: 10.1145/505532.505535.

[44] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, “Selecting Empirical Methods for Software Engineering Research,” in Guide to Advanced Empirical Software Engineering, F. Shull, J. Singer, and D. I. K. Sjøberg, Eds. London: Springer London, 2008, pp. 285–311. doi: 10.1007/978-1-84800-044-5_11.

[45] M. Höst, B. Regnell, and C. Wohlin, “Using Students as Subjects—A Comparative Study of Students and Professionals in Lead-Time Impact Assessment,” Empirical Software Engineering, vol. 5, no. 3, pp. 201–214, Nov. 2000, doi: 10.1023/A:1026586415054.

[46] M. Svahnberg, A. Aurum, and C. Wohlin, “Using students as subjects - an empirical evaluation,” in Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement - ESEM ’08, Kaiserslautern, Germany, 2008, pp. 288–290. doi: 10.1145/1414004.1414055.

[47] “Cronbach’s Alpha.” https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/cronbachs-alpha-spss/ (accessed Jun. 19, 2021).