Validation of Object Oriented Metrics for Evaluating Understandability of Data Warehouse Models

Jaspreeti Singh
Assistant Prof., CSE Dept
USIT, GGSIPU, Dwarka
Delhi-75

Srishti Vashishtha
Student, M.Tech CSE Dept.
USICT, GGSIPU, Dwarka
Delhi-75

ABSTRACT
Datawarehouse has a key role in formulating strategic decisions thus it is very essential to maintain its quality. Metrics have been generally used to direct designers to develop quality data models. Numerous researchers have proposed metrics for multidimensional models for datawarehouse. These metrics are required to be empirically validated to prove their practical utility. Empirical validation of the object oriented metrics for multidimensional models for data warehouses at a conceptual level is presented in the paper. Quality attribute understandability is assessed through various combinations of metrics. Univariate and Multiple linear regression analysis have been used in this paper for computing the multidimensional models quality. The results show that these metrics may be considered as key indicators for quality of multidimensional data models.

Keywords:
Datawarehouse, Metrics, Multidimensional models, Quality attributes

1. INTRODUCTION
In today's world organizations need to gather, store and process huge volume of data that is needed to perform day to day operations. They are able to do this at a comparatively low cost but fail to provide quality information [6]. Inmon provided the solution of adopting datawarehouse, which is defined as "collection of subject-oriented, integrated, non-volatile data that supports the management decision process '' [4]. Datawarehouse is a type of environment for the purpose of providing strategic information for analysing, discerning trends and monitoring performance. It is essential for the organizations to ascertain the quality of the information that they are getting from the datawarehouse. The information quality in a datawarehouse is determined by the data quality, presentation quality and the quality of data model (conceptual, logical and physical data model) [3]. Our aim in this paper is to ensure quality by evaluating understandability of multidimensional models.

Structural properties have been recognised as major factors influencing quality of a software product. Metrics based on structural properties have been widely used to assess the quality attributes like understandability, maintainability; fault-proneness etc. of a software artefact [8].Our present focus is on understandability as it is the key measure of quality of datawarehouse conceptual models. There is a relationship between structural properties, cognitive complexity, understandability and external quality attributes. Structural complexity affects cognitive complexity it in turn affects analyzability, understandability & modifiability; and these

further affect external quality [15]. Hence, the structural complexity plays a vital role while assessing the quality of a model. These complexity metrics helps to estimate the quality of information provided by the datawarehouse. A slight error in the information can cause huge losses to organizations, thus it is important to maintain the quality.
Researchers have recommended quality attributes for multidimensional models and have also proposed metrics to estimate these quality attributes. In this paper we focus on the quality attribute - understandability of multidimensional models. Metrics are the objective indicators of quality. They provide a way of measuring quality factors in a consistent and objective manner. Metrics could be useful to understand and improve software development and maintenance of projects and to maintain the quality of a system highlighting the key problematic areas. A set of metrics for data warehouse models is already presented to compute the structural complexity of a multidimensional model by Serrano et al. [15]. The author has suggested that these metrics need to be validated to ensure the practical utility of these metrics and to draw a final conclusion which may be applied in practice. Even though several quality frameworks for data models have been proposed, most of them lack valid quantitative measures to calculate the quality of conceptual data models in an objective way. This family of experiments is a significant aspect in the process of validating metrics as it is extensively accepted that only after executing a family of experiments; it is possible to develop the collective knowledge to extract constructive measurement conclusions to be applied in practice. [15] [16] Hence, in this paper, we have executed empirical validation by considering a dataset consisting of eighteen multidimensional schemas for a datawarehouse using correlation and linear regression.

The rest of this paper is structured as follows: Section 2 explains the process of metrics validation. Section 3 discusses the metrics of the object oriented models for data warehouses. Section 4 elaborates the experimental setup. Section 5 shows the result of various analysis methodologies. Sections 6 discuss about the threats to the validity of the results and limitations of the study. In the end, Section 7 summarizes the work and presents the conclusion.

2. METRICS VALIDATION PROCESS
Metrics Validation process has certain steps to ensure the reliability of the proposed metrics. It is necessary to follow these steps. Figure 1 presents the method we follow for the metrics proposal [3].

In this figure we have three central activities:

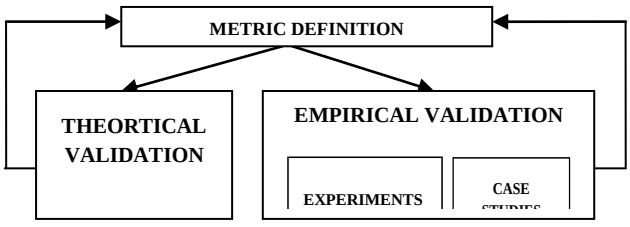

Figure 1. Steps followed in the definition and validation of metrics ref [3]

Metrics definition The initial step is the proposal of metrics. This definition is in accordance to the specific characteristics of the system we want to evaluate and the experience of designers of these systems. [3]

Theoretical validation The next step is the formal validation of the metrics. The formal validation assists us to know when and how to apply the metrics. There are two major tendencies in metrics validation: the frameworks based on axiomatic approaches [17] [2] and the ones based on the measurement theory [18]. The power of measurement theory is the formulation of empirical conditions from which we can derive hypothesis of reality. The final information when applying these kind of frameworks is to know to which scale a metric pertains and based on this information we can identify which statistics and which transformations can be performed on the metric.

Empirical validation. The objective of this step is to prove the practical utility of the proposed metrics. Although there are many ways of performing this step, we can divide the empirical validation into experimentation and case studies [1] [3].

As shown in Fig. 1, the process of defining and validating metrics is evolutionary and iterative. Metrics could be redefined or discarded based on the feedback from theoretical or empirical validations. The aim behind the empirical validation is to prove the practical utility of the proposed metrics. Thus empirical validation is vital for the success of any project.

3. METRICS FOR MULTIDIMENSIONAL MODEL

The metrics proposed by Serrano et al. [15] for the star level is composed of a fact class together with all the dimension classes and associated base classes. The proposed metrics are simple, which is a desirable property of software metrics [5]. These metrics focus on the star level metrics of star schema of data warehouse conceptual model. The metrics are as follows [15]:

1. **NDC(S)** :Number of dimension classes of the star S (equal to the number of aggregation relationships)
2. **NBC(S)** : Number of base classes of the star S
3. **NC(S)** : Total number of classes of the star S
   NC(S) = NDC(S) + NBC(S) + 1
4. **RBC(S)**: Ratio of base classes. Number of base classes per dimension class of star S
5. **NAFC(S)** : Number of FA attributes of the fact class of the star S NADC(S): Number of D and DA attributes of the dimension classes of the star S
6. NABC(S) : Number of D and DA attributes of the base classes of the star S
7. NA(S) Total number of FA, D and DA attributes of the star S
   NA(S) = NAFC(S) + NADC(S) + NABC(S)
8. NH(S) : Number of hierarchy relationships of the star S
9. DHP(S) : Maximum depth of the hierarchy relationships of the star S
10. RSA(S): Ratio of attributes of the star S. Number of attributes FA divided by the number of D and DA attributes, where FA is 'Fact Attribute': Attributes of this stereotype represent attributes of Fact classes in a MD model. D is 'Descriptor Attribute': Attributes of this stereotype represent descriptor attributes of dimension or base classes in a MD model. DA is 'Dimension Attribute': Attributes of this stereotype represent attributes of dimension or base classes in a MD model.

4. EXPERIMENTAL SETUP

4.1 Aim

The goal of the experiment is to analyze the metrics of multidimensional models for the purpose of computing them with respect to the data warehouse quality attributes i.e. understandability.

4.2 Schemas

Eighteen multidimensional data warehouse schemas were used for performing the experiment. Schemas with different metrics values are considered. Metrics values for all the eighteen schemas are given in Table1. All the schemas were taken from different domains.

4.3 Subjects

Total of twenty five M.Tech. students from Guru Gobind Singh Indraprastha University, New Delhi participated in the experiment as subjects.

4.4 Hypothesis

Null hypothesis $H_{0:}$ There is no noteworthy correlation between the metrics and the understandability of the datawarehouse conceptual models.

Alternative hypothesis H_{01}: ¬ H_{11} There is a significant relation between the metrics and the understandability of the data warehouse conceptual models.

4.5 Collected Data

In this paper we have considered one dependent variable, Understandability: It is measured as the time taken by each subject to perform the tasks of the experimental test. The independent variables used in this experiment are the metrics, those variables for which the effect should be calculated. The experimental task consists of understanding a schema and answering the questions based on the schema. Understanding time varies according to the complexity of the schemas and understandability level of the subject. Table 2 represents the collected data (understanding time) of all the subjects for each schema.

Table 1 Values of metrics for the schemas

	NDC	NBC	NC	RBC	NAFC	NADC	NABC	NA	NH	DHP	RSA	Understanding time
SC01	4	9	2.25	3	11	15	29	4	3	3	0.11	65.76
SC02	5	17	22	3.4	7	22	26	55	5	4	0.14	79.52
SC03	6	9	15	1.5	5	18	23	46	4	2	0.12	74.60

SC04	7	15	22	2.14	5	30	19	54	3	3	0.10	79.40
SC05	6	9	15	1.5	4	17	23	44	4	3	0.10	73.20
SC06	3	10	13	3.33	4	9	8	21	2	2	0.23	63.10
SC07	5	8	13	1.3	3	23	20	46	4	3	0.06	64.60
SC08	4	9	13	2.25	4	20	17	41	3	3	0.10	62.40
SC09	6	13	19	2.16	5	26	24	55	4	5	0.10	78.30
SC10	4	13	17	3.25	5	19	17	41	3	3	0.13	76.80
SC11	3	8	11	2.66	4	13	8	25	2	2	0.19	63.24
SC12	5	15	20	3	3	24	20	47	4	5	0.06	79.00
SC13	4	4	8	1	1	23	12	36	2	3	0.02	56.18
SC14	4	9	13	2.25	1	20	11	32	2	3	0.03	64.44
SC15	6	14	20	3.33	5	18	26	47	4	5	0.11	77.00
SC16	8	14	22	1.4	10	30	12	52	6	4	0.23	80.00
SC17	4	6	10	1.5	3	15	6	24	3	2	0.14	58.00
SC18	5	7	12	1.4	4	20	8	32	4	2	0.14	64.80

Table 2 Understanding time (in seconds) of schemas for each subject

	Sub1	**Sub2**	**Sub3**	**Sub4**	**Sub5**	**Sub6**	**Sub7**	**Sub8**	**Sub9**	**Sub10**	**Sub11**	**Sub12**	**Sub13**
SC01	65	55	59	72	79	63	69	71	61	64	67	66	65
SC02	58	68	76	100	96	78	80	74	81	79	99	74	78
SC03	90	61	93	66	60	73	79	70	81	75	80	71	69
SC04	81	69	77	91	71	80	75	86	69	72	87	78	76
SC05	60	48	54	61	69	57	72	70	64	61	79	68	74
SC06	52	56	65	85	53	62	68	57	49	66	72	65	63
SC07	64	71	69	57	79	59	63	66	72	48	60	57	53
SC08	62	72	55	69	74	60	59	65	59	54	61	63	59
SC09	66	76	72	78	94	77	91	86	71	85	80	91	82
SC10	66	95	63	87	94	78	73	86	72	77	79	64	71
SC11	78	50	77	60	57	63	61	71	57	74	63	55	51
SC12	82	73	95	69	64	84	89	66	65	83	87	94	76
SC13	65	54	51	72	46	53	57	56	50	60	52	49	57
SC14	67	58	63	78	58	65	69	57	60	75	64	60	70
SC15	76	83	71	79	72	69	78	65	70	79	81	73	68
SC16	84	78	92	67	70	65	82	78	87	68	75	86	77
SC17	50	59	52	78	51	54	61	54	62	52	61	66	59
SC18	67	55	72	68	64	77	69	54	63	59	61	58	54

	Sub14	**Sub15**	**Sub16**	**Sub17**	**Sub18**	**Sub19**	**Sub20**	**Sub21**	**Sub22**	**Sub23**	**Sub24**	**Sub25**
SC01	64	69	61	70	64	63	68	66	61	69	70	63
SC02	76	85	71	92	82	74	79	69	80	65	95	79
SC03	71	76	82	71	82	67	90	66	72	69	80	71
SC04	81	73	95	82	72	86	73	96	76	78	71	90
SC05	77	86	90	72	89	79	75	92	70	89	78	96
SC06	64	68	57	76	60	51	63	57	62	69	61	75
SC07	59	61	74	63	68	50	65	80	61	59	77	81
SC08	54	65	63	52	64	56	65	75	62	71	58	65
SC09	61	72	80	78	91	67	76	65	80	73	85	81
SC10	68	66	70	74	80	77	65	80	87	69	97	84
SC11	72	64	51	59	66	67	75	59	55	61	66	69
SC12	62	89	77	85	88	80	89	74	65	93	68	79
SC13	69	50	52	53	59	60	55	46	63	57	58	60
SC14	77	55	62	71	78	58	65	69	57	55	65	55
SC15	70	97	77	69	85	94	72	79	82	77	87	73

SC16	61	87	77	89	90	76	81	87	85	95	87	76
SC17	63	50	53	58	49	53	61	65	52	67	63	57
SC18	61	64	71	58	57	69	78	66	70	67	77	61

5. ANALYSIS AND INTEPRETATION

In this section, we have provided the information about the analysis performed to find the relationship between metrics and quality attributes given. We employed both the univariate i.e. linear and multiple regression analysis. The univariate/linear analysis is used to find the ''best'' line to fit two variables so that one attribute can be used to figure out the other. Multiple linear regression is an extension of linear regression, which allows a response variable to be modelled as a linear function of two or more predictor variables

5.1 Descriptive Statistics

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a purposeful way. Descriptive statistics are used to outline the basic features of the data in a study [7]. It provides summary of the data. The descriptive statistics of metrics are shown in Table3. The table shows the ''mean'', ''standard deviation'' ''min'', ''max'' for all metrics considered in this study.

Table 3 Descriptive statistics

Metrics	Mean	Standard Deviation	Min	Max
NDC	4.94	1.349	3.00	8.00
NBC	10.50	3.585	4.00	17.0
NC	15.44	4.422	8.00	22.0
RBC	2.20	0.802	1.00	3.40
NAFC	4.22	2.045	1.00	10.0
NADC	19.88	5.779	9.00	30.0
NABC	16.39	6.652	6.00	26.0
NA	40.38	11.067	21.0	55.0
NH	3.50	1.098	2.00	6.00
DHP	3.16	1.043	2.00	5.00
RSA	.117	0.0579	.02	.23

5.2 Correlation Analysis

We applied Spearman's rho correlation to test the correlation between the independent (metrics) and dependent variable at the significance level=0.05 (widely used value). We evaluated the correlation among each metric with understanding time, which is an important static quantity as shown in Table4. From these results we found that understandability is correlated with NDC, NBC, NC,

NAFC, NADC , NABC, NA, NH, DHP. Significance value greater than 0.5 are highlighted in bold in Table4. The metrics RBC and RSA are not related with corresponding understanding time

Table 4 Spearman's rho correlation

Metrics	**Understanding Time(Correlation)**
NDC	.000
NBC	.000
NC	.000
RBC	**.349**
NAFC	.001
NADC	.029
NABC	.006
NA	.000
NH	.001
DHP	.004
RSA	**.080**

5.3 Univariate regression Analysis

Univariate analysis explores each variable in a dataset separately. The analysis is carried out with the description of a single variable in terms of the applicable unit of analysis. It is concerned with the description or summarization of individual variables in a given dataset [7]. This analysis is done to examine the effect of individual metric on the dependent variable.

The results in Table 5 indicate that there is a significant relationship between NDC, NBC, NC, NAFC, NADC, NABC, NA, NH, DHP and understanding time .They are at significance level=0.05. The metrics having values more than significance level i.e. 0.05 are highlighted in bold in Table5. The metrics RBC and RSA are not significantly related with understanding time. From Table 5 we can observe NC metric is the most significant metric amongst all with the highest value of F values i.e. 155.266. Metrics in decreasing order of F values are as follows: NC NBC NA NDC NABC NAFC NADC DHP NH RBC and RSA.

Table 5 Univariate analysis of the metrics with understandability time as dependent variable

Model		**Sum of Square**	**Degree of freedom**	**Mean square**	**F**	**Sig.**
NDC	Regression	594.722	1			
	Residual	551.943	16	594.722	17.240	.001
	Total	1146.665	17	34.496		
NBC	Regression	935.899	1			
	Residual	210.766	16	935.899	71.047	.000
	Total	1146.665	17	13.173		
NC	Regression	1039.542	1			
	Residual	107.124	16	1039.542	155.266	.000
	Total	1146.665	17	6.695		

RBC	Regression	157.472	1			
	Residual	989.193	16	157.472	2.547	**.130**
	Total	1146.665	17	61.825		
NAFC	Regression	536.727	1			
	Residual	609.938	16	537.727	14.080	.002
	Total	1146.665	17	38.121		
NADC	Regression	315.136	1			
	Residual	832.529	16	315.136	6.064	.026
	Total	1146.665	17	51.971		
NABC	Regression	554.853	1			
	Residual	591.812	16	554.853	15.001	.001
	Total	1146.665	17	36.988		
NA	Regression	750.942	1			
	Residual	395.723	16	750.942	30.362	.000
	Total	1146.665	17	24.733		
NH	Regression	479.741	1			
	Residual	666.924	16	479.741	11.509	.004
	Total	1146.665	17	41.683		
DHP	Regression	480.913	1			
	Residual	665.752	16	480.913	11.558	.004
	Total	1146.665	17	41.610		
RSA	Regression	20.414	1			
	Residual	1126.251	16	20.414	.290	**.598**
	Total	1146.665	17	70.391		

Table 6 Multiple regressions for combination of two independent variables with understandability time

Model	Sum of Square	Degree of freedom	Mean square	F	Sig.
NBC,NC					
Regression	661.030	2			
Residual	300.262	15	330.515	16.511	.000
Total	961.292	17	20.017		
NBC,NA					
Regression	695.166	2			
Residual	266.126	15	347.583	19.591	.000
Total	961.292	17	17.742		
NBC,NDC					
Regression	661.030	2			
Residual	300.262	15	330.515	16.511	.000
Total	961.292	17	20.017		
NBC,NABC					
Regression	672.195	2			
Residual	289.096	15	336.098	17.439	.000
Total	961.292	17	19.273		
NBC,NAFC					
Regression	631.229	2			
Residual	330.063	15	315.614	14.343	.000
Total	961.292	17	22.004		
NBC,NADC					
Regression	656.557	2			
Residual	304.735	15	328.278	16.159	.000
Total	961.292	17	20.316		
NBC,DHP					
Regression	630.314	2			
Residual	330.978	15	315.157	14.283	.000
Total	961.292	17	22.065		
NBC,NH					
Regression	657.076	2			
Residual	304.216	15	328.538	16.199	.000
Total	961.292	17	20.281		

NBC,RBC					
Regression	699.067	2			
Residual	262.224	15	349.534	19.994	.000
Total	961.292	17	17.482		
NBC,RSA					
Regression	655.784	2			
Residual	305.507	15	327.892	16.099	.000
Total	961.292	17	20.367		

Table 7 Multiple regressions for understandability time

Model	Sum of Square	Degree of freedom	Mean square	F	Sig.
NC					
Regression	1039.542	1	1039.542	155.266	.000
Residual	107.124	16	6.695		
Total	1146.665	17			
NC,NBC					
Regression	1043.306	2			
Residual	103.359	15	521.653	75.705	.000
Total	1146.665	17	6.891		
NC,NBC,NA					
Regression	1046.071	3			
Residual	100.594	14	348.690	48.528	.000
Total	1146.655	17	7.185		
NC,NBC,NA,NDC					
Regression	1046.071	3			
Residual	100.594	14	348.690	48.528	.000
Total	1146.665	17	7.185		
NC,NBC,NA NDC,NABC					
Regression	1062.035	4			
Residual	84.630	13	265.509	40.785	.000
Total	1146.665	17	6.510		
NC,NBC,NA,NDC,NABC,NAFC					
Regression	1063.133	5			
Residual	83.532	12	212.627	30.545	.000
Total	1146.665	17	6.961		
NC,NBC,NA,NDC,NABC,NAFC,NADC					
Regression	1075.727	6			
Residual	70.938	11	179.288	27.801	.000
Total	1146.665	17	6.449		
NC,NBC,NA,NDC,NABC,NAFC,NADC,DHP					
Regression	1075.728	7			
Residual	70.937	10	153.675	21.664	.000
Total	1146.665	17	7.094		
NC,NBC,NA,NDC,NABC,NAFC,NADC,DHP,NH					
Regression	1077.440	8			
Residual	69.225	9	134.680	17.510	.000
Total	1146.665	17	7.692		
NC,NBC,NA,NDC,NABC,NAFC,NADC,DHP,NH,RBC					
Regression	1096.193	9			
Residual	50.472	8	121.799	19.306	0.00
Total	1146.665	17	6.309		
NC,NBC,NA,NDC,NABC,NAFC,NADC,DHP,NH,RBC,RSA					
Regression	1102.189	10			
Residual	44.476	7	110.219	17.347	.001
Total	1146.665	17	6.354		

Table 8 Model summary of understanding time

Model	R	R^2	Adjusted R^2
NC	.952	.907	.901
NC,NBC	.954	.910	.898
NC,NBC,NA	.955	.912	.893
NC,NBC,NA,NDC	.955	.912	.893
NC,NBC,NA,NDC,NABC	.962	.926	.903
NC,NBC,NA,NDC,NABC,NAFC	.963	.927	.897
NC,NBC,NA,NDC,NABC,NAFC,NADC	.969	.938	.904

NC,NBC,NA,NDC,NABC,NAFC,NADC,DHP	.969	.938	.895
NC,NBC,NA,NDC,NABC,NAFC,NADC,DHP,NH	.969	.940	.886
NC,NBC,NA,NDC,NABC,NAFC,NADC,DHP,NH,RBC	.978	.956	.906
NC,NBC,NA,NDC,NABC,NAFC,NADC,DHP,NH,RBC,RSA	.980	.961	.906

5.4 Multiple regressions Analysis

Multiple regressions allow us to predict one response variable using two or more predictor variables. The general motive of multiple regressions is to learn more about the relationship between several independent variables and a dependent variable. This technique looks at the relationship between independent and dependent variables and also considers the former in combinations, in order to best explain the variance of dependent variable due to independent variables and finally derive predictions.

In this section the results of multiple regressions to see the effect of various combinations of independent variables on the dependent variable i.e. understanding are presented. [8]Table 6 represents multiple regressions for combination of two independent variables with understandability time. It is noticed that for each model i.e. every combination of two metrics has noteworthy relation with the output variable i.e. understanding time.

In a similar way, we find the multiple regressions by combining every permutation of metrics. We have displayed the final results of the multiple regressions which are strongly related to understanding time i.e. we have combined the metrics in the descending order of F values given in Table 5. Combination of metrics in the descending order of F values is shown in Table 7.

Now, according to the combination of metrics given in Table 7, we found the value of R, R^2 and adjusted R^2. Here, R is the sample coefficient used for linear regression. R^2 is the coefficient of determination which gives information about the goodness of fit of a model. In regression, the R^2 is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data. Adjusted R^2 is a

modification of R^2 that adjusts for the number of explanatory terms in a model.

Table 8 represents the regression model summary of understanding time. It shows the variance explained due to independent variables by observing value of R^2. We observed the change in variance when these metrics are combined. We started with NC as it explains maximum variance in the dependent variable as compared to other metrics. Table 8 represents the final results. It shows that in the final model value of adjusted R^2 is .906 i.e. 90.6 % of independent variables have been covered, which indicates a considerably good quality model .We have also observed that change in variance explained due to NDC is nil, hence NDC does not have a significant effect on understandability.

6. THREATS

This section discusses about the threats to the validity of results and limitations. As we know, several types of threats to the validity of results of an experiment exist. In this section we talk about threats to construct, internal, and conclusion validity. To assure construct validity, it is necessary to perform more experiments. For internal validity, all the subjects possess same degree of experience in data warehouse and modelling. Subjects have never performed this kind of experiment in past, so in our case persistent effects are not present. For conclusion validity, we have taken a sample value of 18, which may not be enough for statistical test. We may attempt a bigger data set through more experimentation.

In this paper we are focusing on two quality factors, i.e. understandability and efficiency. One of the limitations of this work is the size of data. Though we performed this experiment with more numbers of schemas as compared to earlier conducted study [15], but we still feel that more data is required to deal with conclusion validity regarding the effect of metrics on quality attributes. Another limitation of this work is that we have conducted this experiment by considering students as subjects. Due to difficulty in getting professionals to perform the experiment, this experiment was done by students.

7. CONCLUSION

In this paper, we have carried out empirical validation of metrics using statistical techniques. The systems under study were eighteen schemas. In this paper, initially we observed a correlation between the metrics and understandability and subsequently found the effect of various combinations of metrics using univariate and multiple regression techniques of statistics. The results in Table 8 displays that metrics have a noteworthy effect on the quality attribute -understandability.

As a future work, it is essential to keep on performing empirical validation exercises by considering professional subjects, crafting new experiments with more cases and different values of metrics, running case studies with real data from industrial environment to ensure conclusion validity. These experiments will assist data warehouse users in many processes of data warehousing. Thus will help in improving the overall development process of a datawarehouse.

8. REFERENCES

[1] V.R. Basili, F. Shull and F. Lanubille. Building Knowledge through families of experiments. IEEE Transactions on Software Engineering.No. 4.456-473, July/August, 1999.

[2] L.C. Briand, S. Morasca and V. Basili. Propertybased software engineering measurement. IEEE Transactions on Software Engineering. 22(1).68-85, 1996.

[3] Calero C., Piattini M., Pascual C., Serrano, M.A. (2001), "Towards Data warehouse quality metrics.In 3rd International workshop on design and Management of Data warehouses (DMDW 2001), Interlaken, Switzerland.

[4] Inmon,, W. H. (1997), "Building Data warehouse", John Wiley & sons.

[5] Fenton N. (1994), "Software measurement: a necessary scientific basis," IEEE Transactions on Software Engineering, Vol. 20, 1994, pp. 199-206.

[6] GARDNER, S.R.: 'Building the data warehouse', Comnzun. ACM, September 1998,41, (9), pp. 52-60.

[7] Gupta SL, Kumar V (2011) Statistical mechanics. Pragati prakashan, Meeru [GS11] Gosain A, Nagpal S, Sabharwal S, "Quality Metrics for Conceptual Models

for Data Warehouse focusing on Dimension Hierarchies'' July 2011 ACM SIGSOFT.

[8] Gosain A, Mann S (2013) ''Empirical validation of metrics for object oriented multidimensional model for data warehouse'', Springer Int J Syst Assur Eng Manag.

[9] Gosain A , Nagpal S , Sabharwal S, Validating dimension hierarchy metrics for the understandability of multidimensional models for data warehouse.

[10] R. Kimball, L. Reeves, M. Ross and W.Thornthwaite. The Data Warehouse LifecycleToolkit, John Wiley and Sons, 1998.

[11] Kumar M, Gosain A, Singh Y, Empirical validation of structural metrics for predicting understandability of conceptual schemas for data warehouse.

[12] Serrano M., Calero C., Piattini M. (2002), "Validating metrics for data warehouses", IEE Proceedings SOFTWARE 149, 161–166.

[13] Serrano M, Calero M, Piattini M (2003) ''Experimental Validation of Multidimensional Data Models Metrics'', IEEE Proceedings of the 36th Hawaii International Conference on System Sciences – 2003.

[14] Serrano MA, Calero C, Trujillo J, Lujan S, Piattini M (2004) ''Empirical validation of metrics for conceptual models of data warehouse.'' Lecture Notes Comput Sci 3084:506–520 (Caise2004).

[15] Serrano M., Trujillo j, Calero C., Piattini M. (2007), "Metrics for data warehouse conceptual models understandability",Journal of Information and Software Technology49, 851-870.

[16] Serrano M, Calero C, Sahraoui HA, Piattini M (2008) ''Empirical studies to assess the understandability of data warehouse schemas using structural metrics.'' Softw Qual J Springer 16:79–106.

[17] E.J. Weyuker. Evaluating software complexity measures. IEEE Transactions on Software Engineering. 14(9). 1357-1365, 1988.

[18] H. Zuse. A Framework of SoftwareMeasurement. Walter de Gruyter, 1998.