Metrics based software supplier selection

Best practice used in the largest Dutch telecom company

Hans Kuijpers
Program Assurance Metrics Desk
KPN
The Hague, the Netherlands
hans.tm.kuijpers@kpn.com

Harold van Heeringen
Sizing Estimating and Control
Sogeti Nederland B.V.
Vianen, the Netherlands
harold.van.heeringen@sogeti.nl

***Abstract*—This article provides insight into a 'best practice' used for the selection of software suppliers at the largest Dutch telecom operator, KPN[1]. It explains the metrics rationale applied by KPN when selecting only one preferred supplier (system integrator) per domain instead of the various suppliers that were previously active in each domain. Presently (Q2 2012) the selection and contracting process is entering its final phase. In this paper, the model that was built and used to assess the productivity of the various suppliers and the results of the supplier selection process are discussed. In addition, a number of lessons learned and recommendations are shared.**

Keywords; Supplier Selection; Outsourcing; Productivity; Request for Proposal; Metrics; Measurement; Quality;

I. INTRODUCTION

This article illustrates a best practice for selecting a software supplier based on productivity metrics. The largest Dutch telecom operator KPN has decided to consolidate software supplier(s) within various domains (being a logical collection of applications and application chains) in order to reach a more mature Managed Service level within the domain. Reducing the number of suppliers should reduce discussions on responsibilities, internal overhead and last but not least external cost. Therefore for three separate domains a Request for Information (RFI) was sent out. This paper focuses on a specific domain for which a procedure was started to reduce the number of suppliers step by step. From halfway the Request for Proposal (RFP) phase, productivity metrics were used as one of the main selection criteria as part of the RFP Finance stream, adding objective elements next to the pricing attributes and 'subjective' elements from the commercial proposals. The main reason for this was the goal to establish output-based pricing based on Function Points. Therefore the suppliers should a) show and prove their actual productivity information b) show how they are planning to establish a viable baseline and c) show how they are going to show productivity improvements by continuous improvement methods. The expectations in using output base pricing are that that software innovation (new developments and enhancements) productivity will be improved significantly, resulting in market average productivity in 2 to 3 years.

Proposals of the suppliers are in general commercial documents mostly in a fancy presentation. A selection team should not be impressed by the appearance of a proposal only, but mostly on the foundation of the arguments in the proposal itself. To avoid too many discussions, objective criteria like productivity figures are used during the selection workshops and the explanation to the dropped out suppliers.

During this RFP phase (February 2011 till December 2011) the number of suppliers was brought back from 5 to 2. The Best And Final Offer (BAFO) phase was in the final decision stage when this paper was send to the IWSM, 15th April 2012. The timeline for this exercise is showed in the next figure.

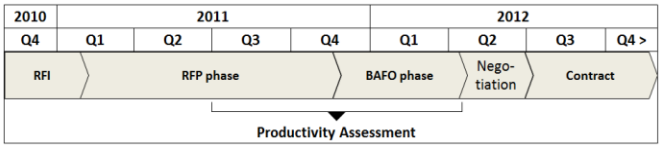

Figure 1. Supplier Selection Timeline

Before comparing the different proposals, the selection criteria should be clear and documented. The KPN Metrics Desk has built an assessment model to be able to rank the submitted project data per supplier in an objective way. The KPN Metrics Desk analyzed the project data and after the analysis, the rank of the suppliers was determined. This paper explains the model, the analysis, the results, conclusions and recommendations.

For confidentiality reasons the supplier names are made anonymous, but the figures are the actual results.

During phase 2 the suppliers were requested to provide insight into their historical projects by sending in a template "Historical Project Data (Appendix A)" per project. The supplier should deliver:

- Six (6) projects, of which a maximum of 3 projects (if possible) need to be performed within KPN domain context;
- Projects should fall within the scope of current technology domain.
- The size of the submitted projects should lay within the range of 300 – 1.000 function points, which is representative for the scope for KPN projects;
- The Functional Sizing Method must be NESMA 2.1 [2] or IFPUG 4.x [3];
- All data fields off the "Historical Project Data" form must be completely filled out;

- Supplier response is limited to the completed template.

In the BAFO phase the suppliers should submit additional reports which show evidence of the size and productivity of the earlier provided projects. Therefore they were requested to release Function Point Analysis reports, Data Collection Forms of the final productivity or they should give insight into their administrative systems.

II. MODEL

KPN wishes to create a long term relationship with its suppliers, based on trust and transparency. Also, KPN wants the supplier to be mature, productive and eager to become more productive over time. With this in mind, the model to assess the productivity of the suppliers was constructed. The most important characteristics to test in this quantitative analysis are:

- The degree of openness and compliancy the suppliers are willing to show;
- The completeness and cohesion of the data submitted. KPN expects the suppliers to act on a CMMI level [6] of at least 3. This means it should not be too hard for them to submit the relevant data;
- The productivity of the suppliers compared to each other and to the 'industry benchmarks';
- The quality of the software the suppliers deliver to their clients for their acceptance test.

The hard part is of course to determine whether the data is representing reality or not. It would not be hard for suppliers to model their data in such a way that they score very high in the model. At this stage, the purpose of the RFP stage was to create a shortlist of two out of five submitting suppliers, so at least they would know that three of them would be gone. However, RFP Finance team decided to trust the data for the time being and to check the data on reality. Unrealistically productive projects are ignored and not analyzed. In the next phase, the BAFO stage, the KPN Metrics Desk is going to deep-dive into the data and audit the correctness of the data submitted.

A. Constructing the model

The model was constructed by the two authors of this paper and was reviewed by RFP Finance team including the Procurement department, Control department and delivery managers of KPN. Based on their input, the weighing factors changed to some degree, but the overall model seemed to be matching the ideas of RFP Finance team very well.

The model tests three criteria per project:

1. The degree of compliancy of the submitted data to the requirements that RFP Finance team stated (Compliancy value: weight factor = 10%)
2. The reality value of a submitted project, compared to objective benchmarks (Reality value, weight factor = 30%)
3. The productivity and quality that the submitted project shows compared to the objective benchmarks (Productivity/Quality value, weight factor = 60%)

The model ranks the suppliers based on the scores on these three criteria and the weight factors mentioned. The detailed behavior of the model is explained in the next paragraphs.

B. Compliancy value

RFP Finance team stated the following requirements with regard to the project data:

1. Supplier should submit 6 completed projects or more. When possible, there should be maximum three relevant KPN projects among the submitted projects. The reason for this is that RFP Finance team wants to compare the KPN projects to the non-KPN projects;
2. The size of the submitted projects should fall between 300 and 1.000 function points;
3. The size must be measured in NESMA 2.1 or IFPUG 4.x function points;
4. All the appropriate data fields in the data collection form should be completely filled in.

The suppliers start with a Compliancy value of 10 points. For every violation of the compliancy, 2 points are subtracted. The number of points a system integrator can get for the compliancy value is therefore 10,8,6,4,2 or 0 points.

The compliancy value counts for 10% in the overall assessment model.

C. Reality value

The project data is stored in a QSM Datamanager [4] file in such a way that the appropriate metrics are calculated by the QSM Datamanager tool. Two of the metrics that the tool calculates are the Productivity Index (PI) [7] and the Project Delivery Rate (PDR) in hours per function point (h/FP). The PI index is a metric that is derived from the QSM SLIM suite. This index is calculated based on the duration, the size and the effort spent on the project. At a given size and a given duration, the more effort spent means a lower PI.

In order to assess the Reality value of the project submitted, the project metrics were compared to two objective benchmarks:

1. The QSM Business trendline based on function points [4]
2. The ISBSG repository 'New developments and enhancements, release 11' [8]

The underlying idea is that RFP Finance team does not believe the project data of projects that are carried out much more

productive than the market average. The criteria to assess whether a project is unrealistic are:

- PI of the project > 'QSM Business PI + 2 standard deviations'

or

- PDR of the project < 'ISBSG P25 PDR'

This means that RFP Finance team does not believe that it is realistic to realize a project with a higher PI than 95% of the projects in the QSM benchmark database. Also RFP Finance team does not trust the data of the projects that are more productive than the 25% best projects in the ISBSG repository (P25), as the ISBSG data is already considered to be best in class data from the industry.

When a project is identified as unrealistic, it is not further analyzed and the project data is discarded from the analysis. It is possible that the assessors override the assessment of unrealistic projects when an explanation is given of the reason why the project performed so very well.

The suppliers start with a Reality value of 10 points. For every unrealistic project, 2 points are subtracted. The number of points a supplier can get for the Reality value is therefore 10,8,6,4,2 or 0 points.

The Reality value counts for 30% in the overall assessment model.

D. Productivity/Quality value

To assess the Productivity/Quality value, the project data is compared to the two aforementioned benchmarks:

1. The QSM Business trendline based on function points
2. The ISBSG repository 'New developments and enhancements, release 11'

The quality is assessed by ranking the median project quality delivered by the suppliers.

1) Productivity assessment (PI score / PDR score)

The Productivity assessment results in a PI score and a PDR score per system integrator.

The Productivity assessment compares the PI realized in the project to the QSM Business trendline at the same size. The absolute distances are measured per project, resulting in the absolute PI delta between the PI of the project and the PI that is considered the market average. Then, the average is calculated by dividing the sum of these absolute distances by the number of projects assessed for that system integrator. This value is called the PI score. The system integrator with the highest PI score is awarded with 10 points, the second gets 8 points, the third gets 6 points, the fourth gets 4 points and the fifth gets 2 points. The PI score weights for 50% in the Productivity/Quality value.

An example of how the PI score is calculated is given in the next figure.

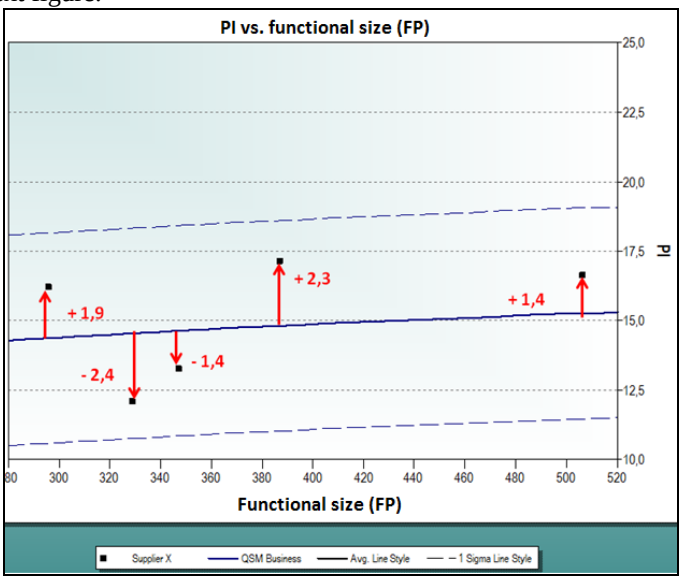

Figure 2. PI Score Calculation

The sum of the distances between the projects (the small rectangles) and the trendline is + 1,8. The average distance to the trendline is +1,8 / 5 = 0,36. If this is the highest score of the five suppliers, 10 points are granted for the PI score.

A similar exercise is done to calculate the PDR score. In this case, the PDR realized is compared to the median PDR of the appropriate ISBSG dataset. The ISBSG median PDR for 3GL projects is calculated by carrying out the following filter criteria on the ISBSG repository R11 [8]

- Data quality A or B (C and D excluded);
- Count approach = IFPUG or NESMA
- Year of delivery > 1999
- Language type = 3GL

This filter results in 221 projects selected. The median of this dataset is 11,4 hours per function point. The ISBSG PDR however comprises the whole project lifecycle (requirements – implementation), while the KPN suppliers submitted data for the 'Construction and Test' phase only. Based on ISBSG experience data, the ISBSG median PDR was lowered with 25% to match on the actual effort spent by the supplier. This means that for 3GL projects, an ISBSG median PDR of 8.6 h/FP was used.

Per project, the absolute difference is calculated, after which the average is calculated (the PDR score). The system integrator with the lowest PDR score is awarded with 10 points, the second gets 8 points, the third gets 6 points, the fourth gets 4 points and the fifth gets 2 points. The PDR score weights for 30% in the Productivity/Quality value.

An example of how the PDR score is calculated is given in the next table:

TABLE I. PDR SCORE

ID	PDR (h/FP)	PDR ISBSG median	PDR score
7	5,9	8,6	-2,7
8	6,0	8,6	-2,6
9	6,9	8,6	-1,7
11	6,2	8,6	-2,4
12	7,3	8,6	-1,3
		Average:	**-2,1**

The projects in this table were all considered to be 3GL projects and therefore the difference between the project PDR and the 3GL ISBSG PDR was calculated. The average difference in this case is minus 2.1. If this is the best value of the five suppliers, 10 points are granted for this.

2) Quality assessment (Quality score)

To assess the quality of the software delivered in the projects submitted by the suppliers, the Defect per 1000 FP metric was calculated for all the projects. This metric only considers the defects that were delivered to the customer (in acceptance test and/or after implementation (1st month of production). This metric was not compared to the ISBSG data, as the ISBSG only collects defect data after implementation.

The median is used, because of the fact that in case of missing defect data, the value of 1.000 is substituted. In case of one project with missing defect data, taking the average value would directly result in a very bad score, while the median could still give a representative figure. This metric is calculated in the following way:

$$\frac{\text{(Defects total – Defects system test)}}{\text{Size (FP).}} * 1.000$$

This means that the model assesses the defects the supplier did not detect in their systems test and therefore the defects that were delivered to their client. For the projects that did not contain data about the defects, a fixed value of 1.000 is put in by the assessors as a penalty for not submitting the defect data. The Quality score is determined by taking the median of the Defect/FP of the suppliers' projects. An example of how the Quality score is calculated is given in the next table.

TABLE II. QUALITY SCORE

ID	Defects/FP	Quality score
15	41,7	
18	13,9	
21	66,7	
22	4,0	
23	10,0	
	Median	**13,9**

The supplier with the highest Quality score is awarded with 10 points, the second gets 8 points, the third gets 6 points, the fourth gets 4 points and the fifth gets 2 points. The quality score weights for 20% in the Productivity/Quality value. The Productivity/Quality value per system integrator is calculated in the following way:

Productivity/Quality value = (Points PI score * 0,5) + (Points PDR score * 0,3) + (Points Quality score * 0,2)

This Productivity/Quality value counts for 60% in the overall assessment model.

E. Total assesment

The total number of points per system integrator is calculated in the following way:

(Compliancy value * 0,1) + (Reality value * 0,3) + (Productivity/Quality value * 0,6).

The system integrator with the highest number of points is ranked first and is the best performing system integrator in this model, based on the project data submitted.

III. RESULTS

A. Processing the data in order to be able to analyze it

KPN provided a project data collection form (appendix A) to the suppliers. The suppliers were expected to send in 6 of those forms, one for each project. Also, the suppliers were expected to fill in all the data fields that KPN requested. The idea was that this should not be too hard for them, as all of them claim to be acting on a high level of maturity and collecting data of completed projects should be part of this maturity. However, the data submitted by the suppliers was not completely what was expected. Some of the findings:

- Some suppliers sent in more than 6 projects, but not compliant to the requirements (at least 3 KPN projects and only projects in the same domain);
- Some suppliers left a number of crucial data fields blank (e.g. defect data, effort data or dates);
- Some suppliers sent in projects that were not finished yet (status: ongoing);
- One supplier sent in a project not measured in NESMA or IFPUG, but in COSMIC[9];
- Some suppliers did not send in different forms, but added data of other projects in the cells where the data of other teams (but on the same project) should be put;
- Some suppliers did not comply to the size requirement and submitted data of either very small projects (< 300 FP), or very big projects (even one of over 5.000 function points!).

The first step was to enter the project data in QSM Datamanager. Only the completed projects measured in NESMA or IFPUG were entered, resulting in one project of

supplier C, two projects of supplier D and two projects of supplier E being excluded. An extract of the remaining project data is given in the next table.

TABLE III. PROJECT DATA

Supplier / Project Name	Size (FP)	PI	PDR Hours/FP	Quality Defects/FP	KPN Project	Generic Domain
Supplier A						
Project A1	365	17,5	4,5	5,5	No	Yes
Project A2	1.567	20,0	5,7	2,6	No	Yes
Project A3	880	23,3	4,5	1,1	No	Yes
Project A4	299	28,0	4,4	1.000,0	No	Yes
Project A5	1.157	17,4	7,0	3,5	No	Yes
Project A6	1.600	17,1	4,0	1,9	No	Yes
Supplier B						
Project B1	96	13,8	5,9	41,7	No	Yes
Project B2	72	16,1	6,0	13,9	No	Yes
Project B3	75	12,9	6,9	66,7	No	Yes
Project B4	855	25,2	7,7	4,7	No	Yes
Project B5	498	19,7	6,2	4,0	No	Yes
Project B6	300	19,3	7,4	10,0	No	Yes
Supplier C						
Project C1	372	15,1	39,0	51,1	No	No
Project C2	397	20,9	19,6	68,0	**Yes**	No
Project C3	342	15,0	20,1	52,6	**Yes**	No
Project C4	317	16,9	19,0	78,9	**Yes**	Yes
Project C5	406	16,8	22,8	1.000,0	No	Yes
Supplier D						
Project D1	5.309	21,4	6,3	5,8	**Yes**	Yes
Project D2	744	17,6	14,4	1.000,0	No	Yes
Project D3	368	16,4	20,9	1.000,0	No	Yes
Project D4	286	17,6	10,4	1.000,0	No	Yes
Project D5	456	18,4	13,2	1.000,0	No	Yes
Project D6	105	13,4	23,8	1.000,0	No	Yes
Supplier E						
Project E1	445	16,3	35,7	222,5	No	No
Project E2	425	16,1	40,8	183,5	No	No
Project E3	601	16,4	44,6	66,6	No	No
Project E4	329	21,2	23,0	340,4	No	No
Project E5	531	23,8	15,2	5,6	No	No
Project E6	506	22,3	18,7	122,5	No	No
Project E7	296	12,8	23,4	54,1	No	No
Project E8	387	17,1	13,2	38,8	No	Yes
Project E9	347	14,1	15,6	46,1	No	Yes

This data forms the basis for the analysis of the model.

B. *Assessing the Compliancy value per supplier*

All the suppliers started with 10 points, and every violation of the requirements resulted in a 2 point subtraction. As it was required that all relevant data fields should be filled in, a blank field resulted already in a two point subtraction. This was done for all the fields that were left empty. The result was that most of the suppliers scored zero points for the Compliancy value, as a lot of relevant data fields were left empty. Other violations to the Compliancy that were identified are:

- Suppliers that did not submit KPN projects (while three were requested). No points were subtracted because it was not known if the supplier has carried out relevant KPN projects in the requested domain at all.
- A number of projects are out of the size range specified (300 FP – 1000 FP). However, the 299 FP project was considered to be so close to the boundary of this range, that no violation was counted in this case.
- Some projects were submitted that were obviously not carried out in the requested domain.
- For some projects, hours and dates were not given per activity (as requested), but aggregated.

The number of compliancy violations turned out to be quite large. Only supplier C managed to get more than zero points. The result of the Compliancy value analysis is given in the next table:

TABLE IV. COMPLIANCY VALUES

Supplier	Compliancy Value
Supplier A	0
Supplier B	0
Supplier C	4
Supplier D	0
Supplier E	0

C. *Assessing the Reality Value per supplier*

1) PI Criterion

One of the two criteria used for assessing the Reality value is the comparison of the project to the QSM Business trendline. This trendline is supplied with the tool, but the underlying data is not known. The trendline is used however as a market average in the business application domain when it comes to the PI metric. In the next figure, the projects of the different suppliers are shown.

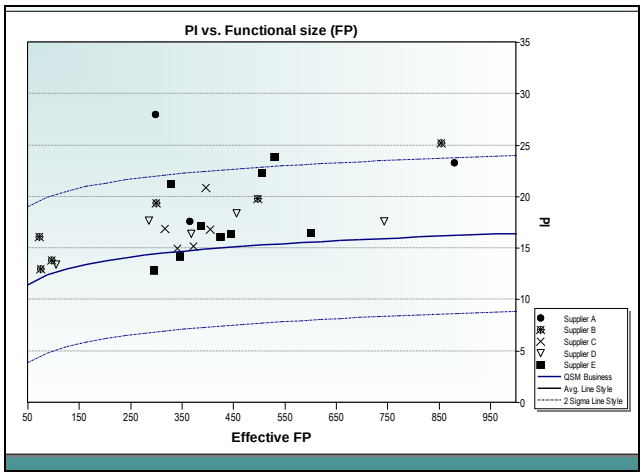

Figure 3. PI Results

The black line is the QSM Business trendline. The dotted lines give the '2 sigma lines', indicating that 95% of the projects that formed the basis for the trendline fall in between these two lines. RFP Finance team believes that projects that were realized against a PI higher than the upper '2 sigma line' are not considered to be realistic. In the figure it becomes evident that there are three projects that are unrealistic: one of supplier A, one of supplier B and one of supplier E. This means that for these three suppliers two points were subtracted from the maximum 10 points based on the PI criterion.

2) PDR Criterion

The second criterion is the PDR criterion. When a project was realized against a better PDR than the P25 of the ISBSG, the project is considered to be unrealistic and discarded for further analysis.

The ISBSG P25 PDR was derived from the ISBSG dataset for two types of projects: 3GL and Legacy (Cobol), as these were the only types of projects submitted by the suppliers. The distinction between these two types is made, because of the fact that in KPN the two types are handled in separate ways in estimating and performance measurement. Both new developments and enhancements are selected, as also in real life both types of projects occur.

The criteria and the results are given in the next table.

TABLE V. DATA SET CRITERIA

Data set	3GL	Legacy
Criteria		
Data Quality	A or B	A or B
Count Approach	NESMA or IFPUG	NESMA or IFPUG
Year of Delivery	> 1999	> 1999
Language Type	= 3GL	-
Primary Progr. Lang.	-	Cobol
Results		
Nr. of projects	221	48
Median	11,4	7,9
Adjusted Median	8,6	5,9
Percentile 25 (P25)	5,4	4,4
Adjusted P25	**4,1**	**3,3**

The table shows that 3GL projects with a PDR smaller than 4,1 hours/FP are considered unrealistic. The same is true for Legacy projects with a smaller PDR than 3,3 hours/FP. There was only one submitted project that was considered to be unrealistic based on the PDR criterion, although not by far. This is project A6 from the table above. This was a 3GL project realized with a PDR of 4,0 hours/FP. There was no explanation submitted by supplier A as to why this project was carried out in such a productive way, and therefore we decided to assess this project as unrealistic and the project was therefore excluded from further analysis.

In the next table, the reality score of the five suppliers is given.

TABLE VI. REALITY SCORE

Supplier	Unrealistic projects PI criterion	Unrealistic projects PDR criterion	Reality Value
Supplier A	1	1	6
Supplier B	1	0	8
Supplier C	0	0	10
Supplier D	0	0	10
Supplier E	1	0	8

D. Assessing the Productivity/Quality value per supplier

The Productivity/Quality value is the most important part of the assessment model. The actual productivity of the suppliers is analyzed and compared to the benchmarks.

For the Productivity/Quality value, the project data is compared to the two aforementioned benchmarks:

1. The QSM Business trendline based on function points (PI score)
2. The ISBSG repository 'New developments and enhancements, release 11' (PDR score)

The quality is assessed by ranking the median project quality delivered by the suppliers to the other suppliers.

Of course, KPN has to take into account that perhaps the suppliers are 'window dressing', which means that they report things more optimistically than they should, and also that suppliers are more likely to send in good projects than the ones that were realized in a less successful way. However, at this stage RFP Finance team is satisfied with the data submitted. In the next stage (the BAFO stage), KPN is going to verify all the data in order to see whether the projects submitted are really carried out as productive as the suppliers wish KPN to believe.

1) PI score

The PI scores are given in the next table.

TABLE VII. PI SCORES

Supplier	PI score	Rank PI score	Points PI score
Supplier A	3,9	2	8
Supplier B	5,0	1	10
Supplier C	3,4	3	6
Supplier D	3,0	5	2
Supplier E	3,2	4	4

Supplier B scored the best on the PI criterion. On average, their projects were realized against a PI that is 5,0 PI points above the trendline. This can be considered as a very high productivity. Because of the model, Supplier B was awarded 10 points for this. All of the suppliers scored an average PI score over more than 0, indicating that they are all on average more productive than the QSM business trendline, which is considered the market average.

2) PDR score

The PDR scores are given in the next table.

TABLE VIII. PDR SCORES

Supplier	PDR score	Rank PDR score	Points PDR score
Supplier A	-3,2	1	10
Supplier B	-2,1	2	8
Supplier C	16,6	4	4
Supplier D	6,2	3	6
Supplier E	18,3	5	2

It turns out that suppliers A and B are again the most productive, but they changed places. Supplier A realized their projects on average 3,2 hours per function point more productive than the ISBSG median and was awarded with 10 points.

3) Quality score

The quality score is given in the next table.

TABLE IX. QUALITY SCORES

Supplier	Quality Score	Rank Quality score	Points Quality score
Supplier A	3,1	1	10
Supplier B	13,9	2	8
Supplier C	52,6	3	6
Supplier D	1000,0	5	2
Supplier E	94,6	4	4

Supplier D only submitted defect data for only one project. For the other projects, the value of 1.000 was substituted, in order to make it possible to analyze the data properly. The median for this supplier was therefore 1.000 defects/FP. Of course, this is not the real quality of their software, but as this supplier apparently does not log defect data, RFP Finance team feels that it is legitimate to punish the supplier for this. Some of the other suppliers also occasionally left out the defect data and they were punished the same way. However, as they still had a number of projects with defect data, the median value is still a 'normal' one.

It turns out that supplier A delivered the lowest number of defects per function point to their clients. Also here, supplier B performed the second best.

4) Overall Productivity/Quality value calculation

The overall Productivity/Quality assessment give the following results:

TABLE X. PRODUCTIVITY/QUALITY VALUES

Supplier	Points PI score	Points PDR score	Points Quality score	Productivity/ Quality value
Supplier A	8	10	10	9,0
Supplier B	10	8	8	9,0
Supplier C	6	4	6	5,4
Supplier D	2	6	2	3,2
Supplier E	4	2	4	3,4
weight	50%	30%	20%	

Suppliers A and B both scored the best on the Productivity/Quality value.

E. The total quantitative assesment.

The total assessment is given in the next table:

TABLE XI. FINAL RANKING

Supplier	Compliancy value	Reality value	Productivity/ Quality value	Total Points	Rank
Supplier A	0	6	9,0	7,2	**2**
Supplier B	0	8	9,0	7,8	**1**
Supplier C	4	10	5,4	6,6	**3**
Supplier D	0	10	3,2	4,9	**4**
Supplier E	0	8	3,4	4,4	**5**
weight	10%	30%	60%		

Suppliers B scored the best in the model, followed by supplier A. Both suppliers did not comply fully with the KPN requirements. Both showed at least five compliancy violations, resulting in a compliancy value of zero. The main difference between the two suppliers is explained by the Reality value. Supplier A submitted two projects that were rejected because they were considered to be unrealistic. Supplier B only submitted one unrealistic project. Supplier C complied best with the KPN requirements and also submitted realistic data. However, their productivity and quality was much lower than suppliers A and B, and because of the weighing factors given, supplier C ended up being third in rank.

IV. FINDINGS DURING BAFO PHASE

After the selection of 2 out of the 5 suppliers as the result of the RFP phase, the BAFO phase started at the end of 2011. Supplier A scored less on other criteria than productivity and therefore suppliers B and C were chosen for the BAFO phase. Unfortunately, the ideas that RFP Finance team had with regard to the validation of the productivity figures could not be carried out completely. It turned out that the suppliers could not share detailed information as this would mean that they would break confidentiality agreements with their customers. For this reason, they could not submit the functional documentation , for KPN to review the function point analysis. Only projects that were carried out for KPN could be validated. However, still the requested insight into the suppliers' administrative systems to check the effort data submitted was not granted. KPN should have made clear beforehand that the project data that the suppliers were going to submit during the RFP phases would be verified in a very detailed way in the BAFO phase. Because this was not made clear beforehand, RFP Finance team did not have the authority to demand this.

Supplier B could not validate the results at all and therefore they provided new project data. RFP Finance team did not like this finding, but accepted the new projects to check the new outcome of the total quantitative assessment. The productivity of these projects was clearly lower than before, which led to a different outcome of the final ranking. This is taken into account during the BAFO evaluation and is registered as a serious issue.

When validating the results of supplier C, a number of FPA measurements showed mismatches to the IFPUG counting guidelines. The provided number of FP's was therefore too low and the PDR too high. The reason for this was that the measurements were carried out by an inexperienced analyst at the supplier side. The supplier was willing to correct the size measurements, but still this issue was also recorded to be discussed in the BAFO selection phase.

The co-operation with the suppliers differs a lot from level to level. Generaly the management was rather closed and reserved, but the people of the suppliers' metrics desks were open and provided as much as possible the requested information when available.

Validation of the projects which were not KPN-related proved to be very difficult, as the suppliers appeal for confidentiality. Therefore KPN asked for conference calls or site visits to their customers to discuss the project data. Supplier B organized a site visit at a Dutch bank and a conference call with a English telecom operator. This proved to be very insightful and gave a lot additional information about the way this supplier is cooperating with their clients when it comes to productivity measurement and improvement.

V. CONCLUSIONS & RECOMMENDATIONS

This productivity assessment was a separate discipline in the total supplier selection and influenced the total outcome significantly. The results ranked the suppliers in productivity performance in a very useful way. The assessment and discussions afterwards gave good insight in the transparency of the supplier and the premises whether they were acting on a CMMI [6] level 3 or higher. The productivity values will be used as starting point for further negotiations with the ultimately chosen partner.

A number of recommendations based on our experiences:

- Make sure the Data collection form (Appendix A) is understood by all the parties involved and that the purpose of this document is clear.
- Make sure that the suppliers understand that the project data will be validated and that they should consult their clients to see whether they may disclose the data before they actually do so. The consequents for violating the governance should be clear for the supplier. For example the supplier will get penalty points, will be discarded for further negotiations or will be replaced by another supplier.
- Construct the model beforehand in an objective way but don't communicate the assessment model itself beforehand to the suppliers. You don't want the suppliers to tweak the data in a way that they score more points. A project start date of two weeks later than actual may for instance already result in a much better PI, perhaps winning 10 points for the PI score instead of 6.
- It is advisable to bring site visits when offered. This gives you the opportunity to collect extra information about suppliers transparency, way of working, contract implementation, etcetera, next to the productivity validation.

REFERENCES

[1] KPN website. http://www.kpn.com

[2] NESMA, Netherlands Software Measurement user Association. "Definitions and counting guidelines for the application of function point analysis, a practical manual." Version 2.1, 2004

[3] IFPUG, International Function Point User Group. Function Point Counting Practises Manual. 4.2. 2004. http://www.ifpug.org

[4] Quantitative Software Management, Software Lifecycle Management toolsuite. http:/www.qsm.com

[5] International Software Benchmarking Standards Group, 'New Developments & Enhacements repository', http:/www.isbsg.org

[6] "CMMI for Development, Version 1.3" (pdf). *CMMI-DEV (Version 1.3, November 2010)*. Carnegie Mellon University Software Engineering Institute. 2010. http://www.sei.cmu.edu/library/abstracts/reports/10tr033.cfm. Retrieved 16 February 2011.

[7] Putnam, Lawrence H. and Myers, W., "Measures for Excellence: Reliable software on time, within budget", NJ: Yourdon Press, 1992.

[8] ISBSG repository "New Developments and Enhancements", version 11, www.isbsg.org.

[9] ISO/IEC19761:2003, Software Engineering "COSMIC, A Functional Size Measurement Method version 2.1, International Organization for Standardization" ISO, Geneva, 2003, www.cosmicon.com.

APPENDIX A: TEMPLATE HISTORICAL DATA[1]

Historical Project

Version 1.0

Project Information

Supplier name:	Supplier X
Project Name:	Project Y
Project ID:	999999
Form submitted by:	Metrics guy
Responsible manager:	Manager
Project Start date:	1-11-2011
Project End date:	20-3-2012

Project Description

	Name	Size (FP)	Defects (systems test)	Defects (total)	Effort (person hours)	(KPN) Application Name
Team 1:	Team 1	378	35	39	8183	Application XYZ
Team 2:						
Team 3:						
Team 4:						
Total:		**378**	**35**	**39**	**8183**	

Time, Effort, Cost, and Staffing

Team 1: Team 1

A1 Phase Name	Start Date (dd/mm/yy)	End Date (dd/mm/yy)	Effort (PHR)
PROJECT MGM.	01-11-11	12-03-12	820
FUNC. DESIGN	01-11-11	14-12-11	740
TECH. DESIGN	14-12-11	25-12-11	514
CODING + UNIT TEST	01-01-12	10-02-12	3624
SYSTEM TEST	10-02-12	24-02-12	2325
TEST MANAGEMENT/QA	14-12-11	12-03-12	120
OTHER TEST	24-02-12	12-03-12	40
Life Cycle	**01-11-11**	**12-03-12**	**8183**

Primary Language	Secondary Language
Cobol	PL/SQL

B1 Non Sizeable activities	Start Date (dd/mm/yy)	End Date (dd/mm/yy)	Effort (PHR)	Description
Data Migration	1-03-12	12-03-12	64	
Implementation	12-03-12	27-03-12	204	
Training				
Environment Maintenance				
Infrastructure				
Total	**1-03-12**	**27-03-12**	**268**	

[1] Data is altered and not corresponding to any of the projects submitted.