

Human Performance Estimating with Analogy and Regression Models: An Empirical Validation

Erik Stensrud
Andersen Consulting
3773 Willow Road, Northbrook, IL
60062,USA
erik.stensrud@ac.com

Ingunn Myrtveit,
Andersen Consulting
3773 Willow Road, Northbrook, IL 60062,USA
&
The Norwegian School of Management
POBox 580, N-1301 Sandvika
ingunn.myrtveit@bi.no

Abstract

Most cost estimation models seem to be validated without testing human performance and using data sets from custom software projects where the software typically is sized in lines of code (SLOC) or function points. From a practitioner's point of view this research seem not to address some important aspects of IT projects that we observe: i) Estimating in an industrial environment is performed by people, not models ii) COTS projects are increasing their market share replacing traditional custom software projects iii) Industrial projects use a large variety of metrics to size the project deliverables and estimate the costs. Estimation by analogy tools like ANGEL [5] and multiple regression analysis provide the necessary flexibility in terms of choice of input parameters. We describe an experiment to evaluate human performance where the subjects were aided by analogy and regression tools respectively. 68 partners and managers in Andersen Consulting estimated 48 different COTS projects. The results in terms of MMRE indicate that users benefit from both tools, however more from regression models than from analogy models as ANGEL. Furthermore, the performance of the ANGEL tool itself is not superior to the performance of the regression model. This result is contradictory to previous studies that claim that ANGEL outperforms multiple regression.

1. Introduction

... In a word, the computer scientist is a *toolsmith* – no more, but no less. It is an honorable calling. If we perceive our role aright, we then see more clearly the proper criterion for success: a toolmaker succeeds as, and only as, the *users* of his tool succeed with his aid. However shining the blade, however jeweled the hilt, however perfect the heft, a sword is tested only by cutting. [2]

Over the years researchers have developed more and more sophisticated estimating models and validated their performance with empirical data. However, from a

practitioner's point of view most of this research seem not to address some important aspects of IT projects that we observe:

- estimating in an industrial environment is performed by *people*, not tools
- industrial projects use a large variety of metrics to size the project deliverables and estimate the costs

Human performance is what counts. The ultimate test of a tool is to evaluate how much it improves human performance whereas the performance of the tool itself is of secondary interest. This because the estimate produced by a model never is the final answer. In real life, practitioners use models and tools as *aids* to make better estimates. Despite this fact, most of the research on project cost estimation seems to draw conclusions on estimating performance based on testing *tool* performance in stead of *human* performance [4]. This would be justified only if it turns out that tools always outperform people using the tools. If tools consistently estimate better than people, we should rely on the estimates produced by the tools and not on human judgment.

Human performance using history constitutes a baseline. Models and tools are not very accurate before they are calibrated with, or have access to, an organization's history with actuals from completed projects. Therefore, we assume that a history must exist in the organization whatever tool is used. Now, a history may be used either by tools or by people without tools. Therefore, a tool adds value only to the extent that the tool improves human performance compared with human performance using the history without the tool. Human performance with history but without tools thus constitute a sort of baseline against which to test the tool.

Tools must be able to handle a variety of size metrics and other metrics. IT projects use many different size metrics because of project idiosyncrasies related to the type of project and the type of technology. A project that

is part of a business transformation initiative uses other product size metrics than a traditional software project, and a COTS project uses different software size metrics than a custom system development project [7]. In addition, most companies already have a history with project actuals from completed projects which they use to estimate new projects. Therefore, we require that estimating models and tools accept a variety of input parameters. Models that require a fixed, predefined set of parameters to produce an estimate are not applicable. An example of a «fixed input» model is COCOMO II that requires either Object Points, Function Points or SLOCs as input [1].

Analogy tools and multiple regression tools accept a variety of size metrics. We have identified two alternative approaches that accept any kind of input parameters and thus provide sufficient flexibility. Apart from expert judgment, these two alternatives are multiple regression analysis and estimation by analogy. The estimation by analogy is a new interesting approach which Shepperd et al. [5] [6] claim outperforms multiple regression models.

Given the observation that estimation tools are to be perceived as *aids* to practitioners, not as replacements of practitioners, and given the observation that only analogy tools and multiple regression tools seem to provide sufficient flexibility to a practitioner who is using a variety of metrics to size and cost his project, the basic questions we want to answer are:

- do analogy tools and multiple regression tools add value to a practitioner? If yes, how much?
- which tool is best, analogy or multiple regression?
- Are tools better than people?

The answers are: «yes», «multiple regression, but not significantly» and «no», respectively.

The remainder of the paper first presents the research questions more formally and then goes on to describe the data and the design of an experiment aimed at testing the research questions. The data analysis and results are presented together with some implications for further research before we conclude.

2. Research Questions and Hypotheses

The research questions are presented formally in table 1. (MMRE is the Mean Magnitude of Relative Error¹).

¹ The estimating performance is evaluated the conventional way, by using MMRE. MRE, the magnitude of relative error is defined as: $MRE = \frac{\text{abs}(\text{ActualValue} - \text{EstimatedValue})}{\text{ActualValue}}$. MMRE, the Mean Magnitude of Relative Error is defined as $MMRE = \text{Average}(MRE_i)$. There are other options to evaluate estimating performance. Some use $PRED(x)=y$. This is a count of the percentage projects (y) that have been estimated with a specified accuracy (x). We did perform $PRED(25)$,

Table 1: Research Hypotheses

	Hypothesis	Formal Hypothesis testing
H1	Having a history, do estimators make better estimates with the additional aid of the output from an estimation by analogy tool?	$MMRE_2 > MMRE_3$
H2	Having a history, do estimators make better estimates with the additional aid of the output from a multiple regression analysis?	$MMRE_4 < MMRE_2$
H3	Do estimators estimate better with the aid of analogy tools than with the aid of multiple regression tools?	$MMRE_4 > MMRE_3$
H4	Do tools estimate better than people who are aided by the same tools? Specifically, does the analogy tool outperform the human estimators aided by the analogy tool? That is, should we rely more on the tool than on human judgment?	$MMRE_A < MMRE_3$
H5	Similar to H4, does the multiple regression tool outperform the human estimators aided by the same tool?	$MMRE_R < MMRE_4$
H6	As for tool performance itself, does the analogy tool outperform multiple regression tools? That is, if we were to rely solely on the tools in stead of on people, is the analogy tool preferable?	$MMRE_R > MMRE_A$

where

- $MMRE_2$ measures human performance with the aid of a history
- $MMRE_3$ measures human performance with the aid of history plus the analogy tool
- $MMRE_4$ measures human performance with the aid of history plus multiple regression models
- $MMRE_A$ measures tool performance of the analogy tool
- $MMRE_R$ measures tool performance of the multiple regression model

3. Data

The data set used for this validation 48 completed COTS² projects extracted from an internal database in Andersen Consulting. The projects span many industries and countries in all regions of the world. The data have been gathered since approximately 1990, and it is an ongoing effort. The data have been reported by project managers who themselves use the database to plan future projects. Therefore, they have an interest in providing

$PRED(50)$ and $PRED(75)$. It did however not change the results, and is not reported.

² All the COTS projects in the sample are of the same type, i.e. it is a very homogeneous data set.

accurate data. In addition, it is a living, dynamic database where information is regularly updated.

Each project has reported data as shown in table 2. There are 10 factors for sizing the product, and effort is reported for three phases. The three phases are Assessment & Planning (AP), Design & Prototyping (DP) and Delivery & Assimilation (DA). Each phase is rigorously defined by a standardized set of activities described in Andersen Consulting’s methodology. In addition to data shown in table 2, the database also lists the COTS modules such as FI (Financials), HR (Human resources), etc. (see table 4).

We obtained the largest data set by having the participants estimate the effort for the Delivery & Assimilation phase in stead of total effort. The DA phase accounts for two thirds of total effort on average which makes it the most important phase in terms of effort consumption.

Table 2: Descriptive statistics for COTS data set³

Variable	N	Mean	Median	StDev	Min	Max
Users	48	346.5	250.0	365.9	7	2000
Sites	48	10.25	4.00	17.72	0	98
Plants	48	7.35	2.00	15.74	0	98
Companies	48	2.833	1.000	5.987	1	35
Interfaces	46	13.07	10.00	10.77	0	50
EDI	35	1.857	0.000	2.830	0	10
Conversions	37	18.38	12.00	18.78	1	93
Modifications	39	9.74	5.00	10.19	0	30
Reports	44	44.16	37.50	32.47	0	100
ModulNo	48	4.500	5.000	2.011	1	8

Effort actuals quality. One may have reasonable confidence in the reported effort data for two reasons. (i) The methodology provides a set of standardized activities and deliverables. This common framework ensures that effort and duration figures across projects are comparable. (ii) Project managers track development time by the hour.

Factor count quality. The COTS project managers reporting the data have received similar training. We may therefore assume a reasonably high interrater⁴ reliability in the counts of factors across projects.

4. Experimental Design

The experiment was done in three parts:

1. history based estimating

2. history plus analogy based estimating
3. history plus multiple regression based estimating

Data collection. The participants completed the three parts in the number sequence given above. Each person estimated the effort for the Delivery & Assimilation phase for the same project three times. Different persons got different projects, randomly assigned, to ensure that MMRE could be computed based on a maximum number of projects. In part one they got information as shown in Table 3 plus a table with 47 projects, i.e. a history with the project to be estimated removed. The history looked similar to table 4 where we have shown a sample history with two projects. In part two, they got the same information as in part one plus the output from the analogy tool, ANGEL, as shown in table 5. In table 5, R1 to R10 is the ranking of the ten closest projects, and EstDA is the estimate produced by ANGEL. In part three, they got the same information as in part one plus the output from a multiple regression model and the model itself (table 6). Also, the participants were explained how ANGEL and regression models worked to let them better judge and use the output.

Time constraints. The experiment required that each person use one hour. We had to make a trade off between making a more realistic, more time consuming experiment on one side and on the other side getting enough experienced participants to permit statistical analysis.

Participant profile. The participants in the experiment have acknowledged skills and at minimum six years of practice. Many of them have 15+ years of relevant practice.

Estimation by analogy. We used ANGEL Lite [5] as the estimation by analogy tool⁵. ANGEL Lite is freeware on the Internet⁶. ANGEL finds the closest project by calculating the Euclidean distance from the project to be estimated to all the other projects in the history. The distance is measured in an optimum subset of the n-dimensional, normalized space. The space is normalized, i.e. all dimensions are in the range 0 to 1, to ensure that all dimensions have equal influence. The tool is also automatically tuned by identifying an optimum subset of the n-dimensional space. For example, in our case five to seven out of the ten dimensions were optimal in most cases. There are several options for finding the optimum subset, among them MMRE and PRED(x). We

³ Effort numbers are considered as sensitive information and are therefore excluded from the descriptive statistics. However, all the projects are industrial projects spanning from 100 to approximately 20.000 workdays.

⁴ Interrater reliability is high if two or more persons get equal counts when counting the same object. For example, if two persons count the number of interfaces in a software system and get equal counts, the interrater reliability of the interface metric is high.

⁵ The ANGEL tool was used in this experiment for convenience and its support for the estimation by analogy approach. Its use and the results of the experiment are neither an endorsement nor a recommendation of ANGEL by Andersen Consulting or the authors.

⁶ http://dec.bournemouth.ac.uk/dec_ind/decind22/web/Angel.html

used MMRE to tune the tool. There are also alternatives for calculating the estimate. ANGEL may compute estimates that are averages or weighted averages of the N closest projects. The simplest, however, is to use the actual value for the closest project as an estimate. We did that because we found that MMRE was lowest using the closest analogy, only. Furthermore, it is trivial for a person to compute averages. We believe that the added value of ANGEL is more in the *ranking* of the closest projects than in the estimate it provides.

One limitation of ANGEL Lite is that all the normalized dimensions have equal weight. For example, the number of users is just as important in normalized space as the number of interfaces in finding the closest analogies. However, there exists a non-free «Deluxe» version of ANGEL that provides the option to weight each dimension.

Best subset linear regression model. We used a best subset linear regression model to estimate the DA effort. The following model was used:

$$DA_Days = 328 + 2.18 \text{ Users} + 554 \text{ EDI} + 101 \text{ Conversions}$$

This is the same as table 6 in equation form. We used a linear model because it is simpler and more intuitive than a non-linear model. Analysis of the residuals did not indicate any non-linearity: The distribution was reasonably normal. The expected value was equal to zero, and the data did not exhibit any particular trends or patterns as a function of the response variable. Furthermore, this is a good model since $R^2(\text{adj})$ is 0.8 indicating that it explains 80% of the variance in effort. Also, all the coefficients are positive as we would assume, and all the coefficients are significant contributors. Since $R^2(\text{adj})$ is 0.8, we have 20% unexplained variance. We believe that the unexplained variance is mainly due to missing independent variables that affect productivity, most notably personnel skill and capability, and probably also reuse.

Table 3: Information provided for project to be estimated

ID	Industry	Users	Sites	Plants	Companies	Interfaces	EDI	Conversions	Modifications	Reports	ModulNo	Modules
151	Manufacturing	1100	7	8	1	25	3	30	24	15	7	FI,CO,AM,MM,PD,SD,HR

Table 4: A sample of the history

ID	Industry	Users	Sites	Plants	Companies	Interfaces	EDI	Conversions	Modifications	Reports	ModulNo	Modules	AP_Days	DP_Days	DA_Days	Total_Days
1	Other	160	2	1	1	5	0	1	0	40	8	AM, BC, CO, FI, MM, PP, PS, SD	200	1130	2296	
2	Manufacturing	320	1	4	1	20	0	30	20	60	7	AM,CO,FI,IM,M,PS,GLX	500	900	3400	4800

Table 5: Output From ANGEL

ID	Best Attributes	MMRE	EstDA	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
151	Sites, Comp, Iface, EDI, Conv, ModulNo	35	3400	2;101		47	136	158	109	48	73	159	155

The R2(adj) metric has some limitations. Does a high R2(adj) mean we have an accurate model? No, the model may still produce inaccurate estimates. The reason is that R2(adj) measures the fit at the mean values. Unfortunately, the predictive ability of the sample regression line falls markedly as the independent variables depart from their mean value. MMRE therefore provides a more realistic measure of a model’s estimating accuracy than R2(adj). Also, MMRE can be used to evaluate other types of models such as the analogy model whereas R2(adj) can only be used to evaluate regression models.

The assumption of normal distribution introduces a small error. Linear regression analysis requires that the dependent variable, effort, be normally distributed. In our case, that is perhaps questionable since the effort will never go below zero but might be high on the upper side i.e. have a long right hand tail. Therefore, the distribution probably is closer to a gamma distribution or a truncated normal distribution.

The assumption of linearity is difficult to validate. In a multi-dimensional space, it is difficult by visual inspection to decide if a model has the right functional form i.e. whether it is linear in the variables or curved. In practice we have to make some judgment and some trial and error.

Table 6: the multiple regression model

	Coef	StDev	T	P
Constant	327.9	490.1	0.67	0.510
Users	2.184	1.076	2.03	0.053
EDI	553.6	111.2	4.98	0.000
Conversi	100.70	24.16	4.17	0.000

$$S = 1696 \quad R^2 = 82.3\% \quad R^2(\text{adj}) = 80.1\%$$

5. Test metrics

The estimating performance is evaluated by using MMRE, and a t-test of mean difference between paired MREs is used to test if the performance differences are statistically significant. Shepperd et al.’s [5] [6] previous

study concluded that ANGEL outperformed multiple regression models based on the difference in MMRE numbers. However, observing differences in means could be caused by chance alone because samples drawn from the same underlying population might have different means.

MRE and MMRE are general as well as reasonable test metrics. Both underestimates and overestimates are to be avoided. Overestimates may lead to premature cancellation of a project due to high implementation costs. Underestimates naturally lead to poor resourcing and results. Therefore, MRE and MMRE seem to be reasonable evaluation metrics.

The T-test of mean difference between paired MREs is a simple test of significance. The t-test compares the means of two groups. The null hypothesis is that the two means are equal. For example, from table 7 we see that $MMRE_A=154\%$ and $MMRE_R=127\%$. So apparently, the multiple regression tool outperforms ANGEL for this data set. However, we have to test if this difference is significant or just random. The t-test tests the significance of the result by creating a single derived variable which is the difference between the paired values and then testing whether this derived mean is zero or if the mean is significantly larger than zero.

MMRE measures accuracy, and SD measures reliability. We need a measure of the reliability, or consistency, in addition to the accuracy to better assess the tools. The standard deviation of MRE (SD_{MRE}) as well as the maximum percentage error (MRE_{max}) are reasonable reliability metrics.

6. Results

Tables 7-9 show human and tool performances. (Figures 1-3 are included to visualize the results in Tables 7-9. The boxplots are interquartile plots, and the line within each box is the median.) Parts 1-3 show human performance when estimating with the aid of a history (part1), an analogy tool (part2) and a multiple regression tool (part3), respectively. The analogy and multiple regression tool performances are shown in the rows ANGEL and MR, respectively.

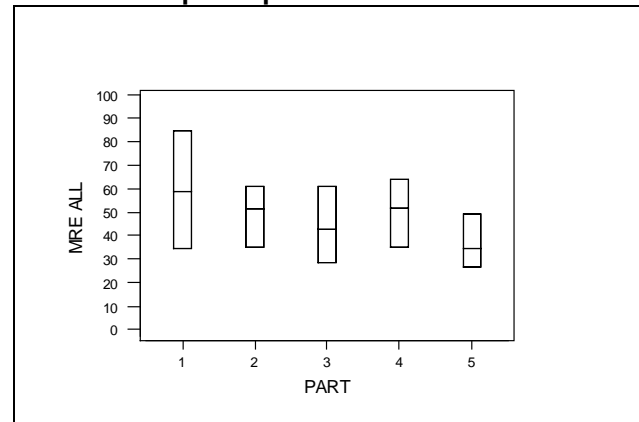
We have divided the participants into two groups based on their experience level. Table 7 shows results for both groups together whereas tables 8 and 9 show results for each group. The first group with approximately 6-9 years experience is named «junior». The other group, i.e. the most experienced people with 9+ years of experience, is named «senior».

Table 7: Estimating Performance Results – All participants

	N	MMRE %	SD_{MRE} %	MRE_{max} %
part 1 – all	61	243	628	3900
Part 2 – all	58	136	230	1208
Part 3-all	56	126	192	900
ANGEL – all	68	154	303	1476
MR –all	68	127	227	1051

N is number of projects, SD is standard deviation of MRE and MRE_{max} is the largest error that was reported in the sample.

Figure 1: Boxplot of Estimating Performance Results – All participants



Both tools improve the average human estimating accuracy by almost a factor of two. The MMRE results in Table 7 suggest that both tools, analogy as well as multiple regression tools, improve the accuracy (from 243% average error down to 136% and 126%, respectively).⁷

Both tools improve reliability by a factor of three or more. The standard deviation (SD) results suggest that both tools improve the estimating reliability by a factor of three (from 628% down to 230% and 192, respectively). The tendency of increased reliability is also supported by the decreasing MRE_{max} (from 3900% down to 1200% and 900, respectively) which suggest that both tools reduce the maximum errors by almost a factor of four. Furthermore, the reliability results suggest that the multiple regression tool aids practitioners somewhat more than the analogy tool does.

Practitioners using the tools perform better than the tools. The MMRE results suggest that practitioners using the analogy tool estimate more accurately than the tool itself (136% vs 154%). Likewise, the standard deviation suggests that practitioners using the analogy tool estimate more consistently and reliably than the analogy tool itself (230% vs 303%). The results exhibit a similar pattern for the multiple regression tool.

Practitioners perform better using the multiple regression tool than when using the analogy tool.

⁷ This trend is consistent across all the evaluation metrics: Mean MRE, Median MRE or PRED(x).

Estimating accuracy is 126% using the regression tool and 136% using the analogy tool. Estimating reliability in terms of standard deviation is 192% using regression and 230% using analogy. The maximum error is 900% using regression and 1208% using analogy.

The multiple regression tool outperforms the analogy tool. Finally, it seems that the multiple regression tool (MR) itself performs better than the analogy tool (ANGEL) in terms of accuracy (127% vs 154%) as well as reliability (standard deviation is 227% vs 303%, maximum error is 1051% vs 1475%).

Table 8 and Figure 2 show a similar overall tendency as Table 7 and Figure 1.

Table 8: Estimating Performance Results – Junior group

	N	MMRE %	SD _{MRE} %	MRE _{max} %
Part1 – jun	40	321	762	3900
Part2 – jun	39	173	269	1208
Part 3 – jun	38	154	217	900
ANGEL – jun	41	177	310	1476
MR- jun	41	169	274	1051

N is number of projects, SD is standard deviation of MRE and MRE_{max} is the largest error that was reported in the sample.

Figure 2: Boxplot of Estimating Performance Results – Junior group

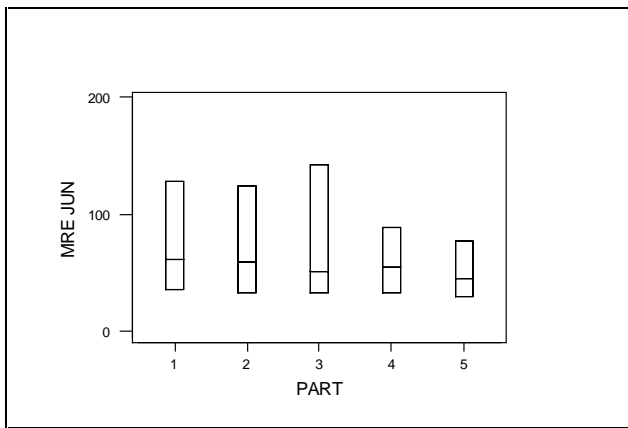


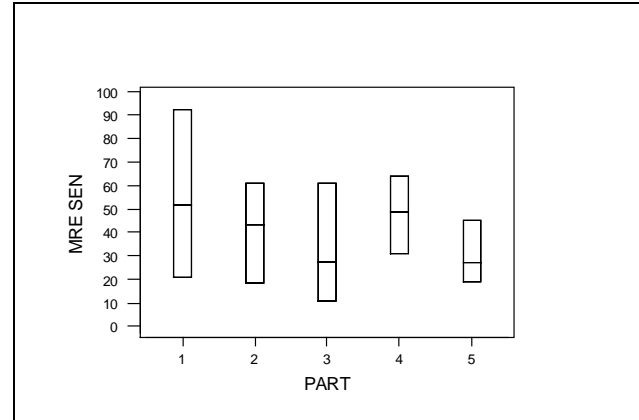
Table 9⁸: Estimating Performance Results – Senior group

	N	MMRE %	SD _{MRE} %	MRE _{max} %
Part1 – sen	21	94	113	355
Part2 – sen	19	60	78	350
Part 3 sen	18	67	107	414
ANGEL-sen	27	119	294	1476
MR-sen	27	64	104	478

⁸ Using the regression model results as the norm, seniors, unintentionally, got easier projects to estimate than juniors. Therefore, any conclusions regarding the differences between the two groups should be treated with caution.

N is number of projects, SD is standard deviation of MRE and MRE_{max} is the largest error that was reported in the sample.

Figure 3: Boxplot of Estimating Performance Results – Senior group



The ANGEL tool is outperformed by the most experienced practitioners and by the multiple regression tool. Table 9 (and Figure 3) show a similar overall tendency as Tables 7 and 8, i.e. that human estimating accuracy and reliability is improved with the aid of both the analogy (ANGEL) and the multiple regression (MR) tools. However, the most important observation is that ANGEL performs worse than anybody else, both seniors without tools, seniors using any of the two tools (including ANGEL), and also compared with the multiple regression tool. ANGEL's worst error (MRE_{max}) is wrong by a factor of 15 whereas seniors are never wrong by more than a factor of four and the regression tool is maximum wrong by a factor of five.

Still, ANGEL adds value to seniors. Despite the fact that ANGEL in some cases performs poorly, senior people using ANGEL perform well suggesting that seniors still benefit from the tool (compared to the results from the junior group in table 8). Senior estimators are more likely to use their expertise to screen the results from ANGEL and discard the extreme errors. It takes human skill to identify when to trust the output and when to discard it as misleading. The more experienced, and probably more confident, practitioners seem to better judge on this matter.

Not surprisingly, seniors estimate more accurately and more reliably than juniors. Comparing Tables 8 and 9, we see that the standard deviation and the largest error are much larger for the junior group. This is most notable when estimating without any tools where we see that the junior group benefited most using the tools. See, however, footnote 8.

In summary, Tables 7-9 suggest a tendency of increasing human performance when aided by tools. The

first two hypotheses are therefore supported by the preliminary findings. The third hypothesis is supported only for the senior group, who seemed to estimate somewhat better with the aid of ANGEL than with the aid of regression models. The preliminary findings suggest that the other hypotheses are rejected: Practitioners using tools estimate better than the tools. As for tool

performance itself, the regression tool seems to estimate more accurately and reliably than the analogy tool.

However, the significance of this tendency in the preliminary findings must be tested. The significance of the results, based on a t-test, are shown in Table 10. This table shows the mean of paired differences with the significance level in parenthesis.

Table 10: Significance of results – t-test of mean of paired differences

	H1 MMRE ₂ > MMRE ₃	H2 MMRE ₂ > MMRE ₄	H3 MMRE ₄ > MMRE ₃	H4 MMRE _A < MMRE ₃	H5 MMRE _R < MMRE ₄	H6 MMRE _R > MMRE _A
All	212 (0.08)*	108,5 (0.07)*	-12.3 (0.74)	12.6 (0.36)	5.9 (0.4)	-27 (0.8)
Senior group	20.7 (0.06)*	1.7 (0.44)	10.8 (0.18)	-12.2 (0.76)	2.28 (0.82)	-55.5 (0.89)
Junior group	133.1 (0.07)*	159 (0.07)*	-23.2 (0.8)	24.7 (0.32)	7.7 (0.41)	-8.2 (0.57)

The asterisks (*) and (**) mean statistically significant at $\alpha=10\%$ and $\alpha=5\%$, respectively.

H1. All practitioners estimate significantly better with the analogy tool than without the tool. This confirms hypothesis H1.

H2. Practitioners estimate significantly better with the multiple regression tool than without the tool. This is confirmed as statistically significant for all but the senior group. Thus, hypothesis H2 is partially confirmed.

H3-H6. None of the other hypotheses are confirmed at a 10%, or better, significance level. Thus, any differences could be caused by chance alone. Specifically, we can not conclude that the analogy tool is superior to the multiple regression tool or vice versa (H3). However, we do see a trend in favor of the regression tool, in particular with respect to reliability. The analogy tool makes a few large errors, but we observe that the most experienced practitioners seem to be able to identify those cases, and therefore they benefit more from this tool than less experienced practitioners. As for H4 and H5, we cannot conclude that practitioners outperform tools or vice versa. As for H6, we can not support Shepperd et al. [5] who claim that ANGEL outperforms multiple regression.

7. Limitations to the results

There is a potential learning effect since each subject estimated the same project thrice, i.e. both with the aid of the dataset, the analogy tool and the multiple regression tool in the same sequential order. However, we believed that the time constraints did not permit much learning from the previous part since in the next part they had to concentrate fully on understanding the introduction and the outputs. Also, the experiment is conservative with respect to human performance for a number of reasons.

Ex post vs ex ante estimating. In this experiment, the participants estimated *ex post*. In practice, estimates are

done *ex ante*. If an estimate is within reasonable limits, it is the manager's job to manage to the estimate. Therefore, we may assume that *ex ante* estimates are more accurate than *ex post* estimates.

Time constraints. There was short time to estimate and limited information. In practice, estimates are validated using several approaches. In general, both bottom-up and top-down techniques are used to «sanity check» the estimate. Furthermore, sensitivity analysis is used to assess likely ranges of the estimates. Also, there was short time to investigate the history. In addition, the history was provided in paper format. A table provided in a spreadsheet application would allow the sorting of columns and rows and other manipulations which we presume would result in a better performance using the history. The introduction to the analogy principles was brief. A deeper understanding of the strengths and limitations of the ANGEL tool and letting the users use the ANGEL tool itself would probably have improved human performance. The introduction to regression model was even briefer, however the participants were more familiar with this tool in general.

Limited information. We gave the participants the same information as we gave to ANGEL except that the people got a list of the actual COTS modules in addition to the number of modules. ANGEL could have used this information as categorical variables with values «1» or «0». We believe, however, that this would have decreased the performance of ANGEL Lite since all dimensions have equal weight.

Few seasoned COTS estimators. Few of the participants were actually experienced in estimating COTS projects. The majority comes from the custom solutions, not COTS solutions, side. The few actual COTS experts all estimated within 30% accuracy on

part1, the expert judgment, and outperformed everybody else. However, this result is not statistically significant.

Single person estimating. In practice, an estimate is reviewed and quality assured by at least one other experienced person. Thus, it is a group effort, not a single person effort. The worst cases of human performance would therefore not occur in practice. However, we had to make a trade-off between the realism on one side and on the other side getting a large enough sample to permit statistical analysis.

8. Some additional findings

We also asked the participants whether they perceived that the tools added value or not, and which tool, inclusive history, they would prefer. Table 11 shows that history is preferred to both tools. Table 12 shows that, on average, they did not perceive any added value of the analogy and regression tools when having the history. However, the estimating results clearly indicate that the tools actually add value.

Table 11: Practitioners' tool preferences

	Prefer ANGEL	Prefer MR	History
all	11	13	17

Table 12: Practitioners' confidence in the estimating tools

Do you have:	yes	no	Somewhat
greater confidence in your estimate with history	20	9	13
greater confidence in your estimate with the aid of ANGEL	14	17	11
greater confidence in your estimate with the aid of regression models	14	14	14

We find it interesting that the objective estimating results differ from the participants' perception. We can speculate on why. One explanation may be that many practitioners are reluctant to use tools that are very different from what they are used to. Furthermore, many practitioners consider that estimating is an art, not a science. Therefore, statistical methods are not generally approved. However, we believe that expert intuition is nothing but experience, i.e. an internalized history and internalized statistics.

9. Implications for further research

The results suggest that human performance improves with the aid of an estimation by analogy tool and a multiple regression tool. However, the performance is in general not satisfactory and needs to be improved. There are a few complementary approaches to improving the estimating performance:

- improve the tools

- partition the history
- leverage expert knowledge

The analogy tool should use weighted dimensions. The ANGEL Lite tool seems to provide added value to a practitioner. However, in some cases the tool identifies closest analogies that result in MREs larger than 1000%. This reduces the confidence in the tool. To produce more stable, robust results the dimensions must somehow be weighted. The problem remains as how to weight them. One possible solution is to base the initial weights on expert opinion and then refine by trial and error. Another idea is to use the Z scores from regression analysis as weights. A third option is to build in some sort of learning in the tool so it learns the importance of the various dimensions [6].

The statistical analysis should use distributions that better model the actual distribution. We will investigate using a gamma distribution or a truncated distribution in stead of a normal distribution since we know that the effort variables are not normally distributed.

The history should be partitioned into smaller and more homogeneous groups. The size of the history presumably influences how easy it is for a person to identify patterns in it. A history that is small in terms of the number of factors and the number of projects is easier to understand for a human mind than a large history. Therefore, one should expect an increasingly added value of tools with increasing size of the history. Opposite, for a small history the added value of tools like ANGEL is probably reduced. Also, a large history may be partitioned into smaller, more homogenous groups to enable comparing apples with apples. For example, assuming that country specific idiosyncrasies affect productivity, it would improve estimating performance if the history was partitioned by country.

Expert knowledge must be better leveraged. There seem to be significant differences between those with extensive COTS estimating experience and the rest. Therefore, we are investigating ways to better leverage expert knowledge across the organization.

10. Conclusions

Both analogy tools and multiple regression tools are valuable to a practitioner. All participants estimated significantly better with both tools (even if they did not acknowledge it).

Both tools decrease the gap between the less experienced and the more experienced practitioners which means that the less experienced practitioners benefit more from the tools than the more experienced practitioners.

The average estimating error was reduced from 320% to 160% for juniors and from 90% to 60% for seniors.

We did not find that the analogy tool outperforms the multiple regression tool. Rather, the results suggest the opposite. This finding applies when evaluating human performance with the aid of the tools as well as when evaluating tool performance itself. However, we can not conclude that the analogy *approach* is inferior to the multiple regression approach based on the performance of the ANGEL Lite *tool* since the performance of ANGEL Lite is limited by applying equal weights to all dimensions. Furthermore, for datasets with missing values for one or more observations, the regression model will not be able to estimate these whereas ANGEL still can.

The seniors benefit more from the analogy tool in some respects. Since the analogy tool in some cases performs very poorly, it takes human skill to identify when to trust the output and when to discard it as misleading. The more experienced, and probably more confident, seniors seem to better judge on this matter. Therefore, one cannot state which tool is best without adding the question «for whom?».

Finally, the results suggest that tools alone do not estimate better than people using the same tools. On the contrary, the results for the seniors suggest the opposite. Therefore, one cannot conclude that one estimating tool is better than another based on testing tool performance alone. Taking human performance into account may alter the conclusion. So, we conclude that empirical validation of estimating tools would benefit from focusing on human performance rather than just testing the performance of the tools themselves.

Acknowledgements

The authors are most grateful to all the subjects, our busy colleagues in Andersen Consulting, who volunteered in the experiment. We would also like to thank the numerous knowledge champions in Andersen Consulting's global SAP community who helped us improve the quality of the COTS dataset and the anonymous referee who provided constructive comments. Finally, we would like to thank the partnership in Andersen Consulting who sponsored the research.

References

1. *COCOMO II Model Definition Manual*, Version 1.4, University of Southern California, <http://sunset.usc.edu/COCOMOII/cocomo.html>, 1997.
2. Brooks, F.P. The Computer Scientist as Toolsmith II. *Comm. ACM*. 39, 3 (March 1996), 61-68.

3. *IFPUG Function Point Counting Practices: Manual Release 4.0*, IFPUG, Westerville OH, 1994.
4. Lederer, A.L. and Prasad, J. A Causal Model for Software Cost Estimating Error. *IEEE Trans. Software Eng.* 24, 2 (Feb 1998), 137-148.
5. Shepperd, M.J., Schofield, C. and Kitchenham, B.A. Effort Estimation Using Analogy. *Proc. ICSE 18* (Berlin, March 1996), 170-178.
6. Shepperd, M.J. and Schofield, C. Estimating Software Project Effort Using Analogies. *IEEE Trans. Software Eng.* 23, 12 (Nov 1997), 736-743.
7. Stensrud, E. and Myrtveit, I. The Added Value of Estimation by Analogy – An Industrial Experiment. *Proc. Eur. Software Measurement Conf. - FESMA'98* (Antwerp Belgium, May 1998), Technologisch Instituut vzw, 549-556.