A Controlled Experiment to Assess the Benefits of Estimating with Analogy and Regression Models

Ingunn Myrtveit and Erik Stensrud, *Member*, *IEEE*

Abstract—To have general validity, empirical results must converge. To be credible, an experimental science must understand the limitations and be able to explain the disagreements of empirical results. We describe an experiment to replicate previous studies which claim that estimation by analogy outperforms regression models. In the experiment, 68 experienced practitioners each estimated a project from a dataset of 48 industrial COTS projects. We applied two treatments, an analogy tool and a regression model, and we used the estimating performance when aided by the historical data as the control. We found that our results do not converge with previous results. The reason is that previous studies have used other datasets and partially different data analysis methods, and last but not least, the tools have been validated in isolation from the tool users. This implies that *the results are sensitive to the experimental design*: the characteristics of the dataset, the norms for removing outliers and other data points from the original dataset, the test metrics, significance levels, and the use of human subjects and their level of expertise. Thus, neither our results nor previous results are robust enough to claim any general validity.

Index Terms—Software cost estimation, commercial off-the-shelf (COTS) software projects, multivariate regression analysis, estimation by analogy, human performance, controlled experiment, enterprise resource planning (ERP) systems.

1 Introduction

It is an important part of an experimental science to have convergence of empirical results and explain disagreements. To have general validity, empirical results must be subjected to ruthless scrutiny and still converge. To be credible, an experimental science must understand, and clearly state, the limitations to the results, and where the results do not converge, it must be able to explain why. As in physics, the ultimate ideal is theories with a maximum explanatory power, "natural laws," simply because a theory with more explanatory power is more useful than a theory with less explanatory power. Since this ideal is often not achievable, the next best is to have theories with clearly understood limitations. As in physics, we must know when Newton's law of gravity is not applicable and you have to use Einstein's theories of relativity instead.

Improving project cost estimation is one of the top priorities in many software development organizations. There is a continuous search for better models and tools to aid project managers in their estimating process, in particular for the enterprise resource planning (ERP) market which is gaining a significant share of the total IT market [6]. The ERP market is dominated by COTS vendors like SAP, Baan, Oracle, and PeopleSoft. Unfortunately, many of the existing estimating models cannot be used to estimate these kind of projects because they use function points [3] or source lines of code (SLOC) as the fundamental size metric. COCOMO [1] is an example of this type of model. In our particular COTS dataset with project actuals, none of the projects have reported function point or SLOC counts. The choice of estimating tools were motivated by this fact. Both Estimation by Analogy [4], [5] and regression analysis are promising approaches for improving estimating accuracy and reliability when a history of completed projects exists since these two approaches accept a variety of input parameters for product sizing and productivity adjustment. Therefore, we found the results of Shepperd et al. [4], [5] interesting where they claim that estimation by analogy outperforms regression models. We, therefore, replicated and extended their study to test if their results were valid also in our environment. The most important difference between their environment and ours is that estimating tools are used as aids by experienced practitioners who provide the final estimate. In the case of Shepperd et al., the results were based on testing tool performance alone. We were concerned both with the question "which tool is best" as well as with the question "how good are the tools."

We found that our results generally do not converge with their results. As for "which tool is best," we did not find that human subjects estimate better using analogy than using regression. As for "how good are the tools," we were not able to replicate their high performance levels for analogy. Our overall results suggest a higher estimating inaccuracy. The reasons why perfect replication was not achieved are stated in Section 11, "Discussion." The results are briefly presented in Sections 8, 9, and 10. Only the results that are relevant to discussing the problems of

- *I. Myrtveit is with The Norwegian School of Management, PO Box 580, N-1301 Sandvika, Norway. E-mail: ingunn.myrtveit@bi.no.*
- *E. Stensrud is with Ernst & Young Consulting, PO Box 6834, St. Olavs plass, N-0130 Oslo, Norway. E-mail: erik.stensrud@ey.no.*

Manuscript received 9 July 1998; revised 25 Feb. 1999.
Recommended for acceptance by D. Ross Jeffery.
For information on obtaining reprints of this article, please send e-mail to: tse@computer.org, and reference IEEECS Log Number 109538.

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 judgement?**	$MMRE_A < MMRE_3$
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?**	$MMRE_R > MMRE_A$

TABLE 2
Descriptive Statistics for COTS Dataset*

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 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.*

convergence are presented. For a fuller presentation of the results, we refer the reader to [6], [7].

2 Research Hypotheses

We conjecture that both estimation by analogy and regression models improve human estimating performance compared with using the historical data only. We also try to confirm or reject the claim that analogy outperforms regression. Finally, we test if tools perform better than people aided by tools. The research questions discussed in this paper are presented formally in Table 1. MMRE is the Mean Magnitude of Relative Error 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

Since the purpose of this paper is to present and discuss the methods and techniques rather than the results, we have limited the research hypotheses to the minimum required to demonstrate and discuss the methods. A more complete list of all the actual research hypotheses and results is provided in [6], [7].

3 COTS Dataset

The dataset used for this validation consists of 48 completed COTS[1] projects. The data have been gathered since 1990, and it is an ongoing effort. All the projects are industrial projects spanning from 100 to 20,000 workdays, and there are 10 factors for sizing the product. See Table 2. A more detailed description of the data is provided in [6], [7].

The credibility of the results depend ultimately on the quality of the data. There has been devoted much research effort to get consistent and standardized measures of software size in an attempt to improve estimating accuracy. However, no chain is stronger than the weakest link. We suspect that data quality in general, and the quality of effort actuals in particular, is the weakest link in empirical studies of project cost estimating.

3.1 Effort

One of the major challenges in gathering project actuals is to ensure a reasonable quality of the effort actuals. Effort actuals are frequently of bad quality for several reasons:

- it is unclear whose time is reported
- it is unclear what time is reported

Whose Time. First, the practice for reporting overhead time such as project manager's time and secretary's time varies across organizations and probably also across projects within the same organization. Second, projects

1. All the COTS projects in the sample are SAP R/3, i.e., it is a homogeneous dataset.

frequently involve many personnel categories: the provider, the client as well as third party personnel. Some projects report time for only their own personnel, excluding time from client personnel, while some do not. We have omitted projects from the sample that have not reported time for all personnel categories. Furthermore, the organization has a standard practice for whose overhead time is reported.

What Time. This concerns the scope of the project. The scope is defined in several dimensions:

- project life cycle
- Work Breakdown Structure (WBS)
- project type

Projects start and terminate at different points in the life cycle. For example, one project may develop a requirements specification whereas another project started out with a given specification. Likewise, one project may terminate when the software is system tested whereas another project continues through roll-out and delivers a product in full operation.

Projects deliver a varying range of end products. For example, one project may train the end users before and after the system is put in operation whereas another project just writes a small user manual.

Projects execute different sets of activities. For example, some projects develop a system from scratch, whereas other projects enhance existing systems. The latter project types include activities to assess and understand the existing system. The former project types do not.

In our sample, the projects have reported time per phase, where each phase is rigorously defined by a set of standard activities and deliverables. Furthermore, this standard methodology plus the homogeneity of these COTS projects ensures that these projects deliver the same range of end products. As for the last point, all projects report whether it is a first release or an enhancement release. However, we did include both project types in the sample. Finally, it should be obvious that the scope is not correlated to any known software size measure. Therefore, we cannot base estimates solely on measuring *software* size.

3.2 Product Size

The product size factors account for the large scope of these COTS projects. For example, the training effort is correlated to the number of users, and the deployment effort is correlated to the number of physical locations such as plants and sites. All the factors are based on counts of physical rather than logical units. For example, the number of interfaces is a count of physical interface modules.

Counts based on physical modules presumably has high interrater reliability. This because module counting could be automated in principle. Counting logical units on the other hand requires some interpretation.

Physical module size is reasonably well correlated with effort. Of course, a physical module may in theory include a varying number of logical units depending on the physical design. So at a first glance, this seems to be a less accurate sizing approach. However, the personnel in the organization all receive similar training, and it is a relatively homogeneous culture. This homogeneity ensures that designers design modules in a reasonably similar fashion. Therefore, a module is a reasonably consistent unit of measure.

Counting physical modules is less time consuming than counting logical units. Counting logical units is a manual, time consuming task. None of our project managers find it pays off to put so much time in the software sizing for estimating purposes. We, therefore, have to take more pragmatic and less time consuming approaches than for example function point that requires the counter to read and learn a several-hundred pages manual such as the IFPUG manual [3].

4 The Analogy Tool

We used ANGEL Lite [4] as the estimation by analogy tool. ANGEL Lite is freeware on the Internet.[2] 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 10 dimensions were optimal in most cases. There are several options for finding the optimum subset, among them MMRE and PRED(x). We 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 nonfree "Deluxe" version of ANGEL that provides the option to weight each dimension.

5 The Multiple Regression Model

The COTS community had selected 10 variables that they considered the main cost drivers. Based on expert knowledge and a best subset regression we came up with the final linear model (see Table 3).

In developing a useful regression model for estimating, a number of concerns must be addressed:

- Have we included the right and most important variables?
- Is the formal model correctly specified?
- How good is the model's predictive power for estimating?

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

TABLE 3
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**

$S = 1,696$; $R^2 = 82.3$ percent; $R^2(\text{adj}) = 80.1$ percent

5.1 Including the Right Cost Drivers

The COTS community had selected 10 variables that they considered the main cost drivers. However, after closer examination we could not use them all as independent variables in a regression model. Performing a simple Pearson correlation, we found that some of these variables were highly correlated with each other. We eliminated the variables that were highly correlated with variables we included in the model. Furthermore, we found that some variables were not clearly and consistently enough defined. Based on this knowledge and a best subset regression we came up with the final model.

5.2 Correct Model Specification

5.2.1 Residual Analysis

Estimation theory for linear regression is tied to certain assumptions about the distribution of the residual. The usual assumption for the residuals are that these terms are distributed as independent, normal random variables with mean zero and identical variances. We verified these assumptions with the aid of diagnostic plots. See Fig. 1.

The analysis of the residuals does not indicate any nonlinearity. The distribution is reasonably normal. (See both Fig. 1 and Fig. 2). The expected value is equal to zero, and the data do not exhibit any particular trends or patterns as a function of the response variable. We, therefore, assume that the model is linear. Since we have a multivariate model with three variables, we also plotted each of the variables against the residual to see if any of them had any indication of hidden nonlinearity. This is not reported.

We accept that the distribution is normal from the plot in Fig. 2. A rule of thumb is that skewness will be in the interval $(-0.5, 0.5)$ if the distribution is normal. In our case it is 0.4, which supports the normal assumption. However we do see a little kurtosis, here -0.5 (Perfect normal would be 3). That means we have some "long tails," which we do know is the case.

However, we do observe that we have some heteroscedasticity. This reveals itself in the plot of residuals vs. fits in Fig. 1. That is, the residuals do not have constant variance; rather they increase with the size. The heteroscedasticity is, however, not very pronounced nor is it surprising. This observation supports our intuition that the absolute precision with which we may predict the effort for large projects is less than for small projects; we assume that the absolute estimating error will increase with the size of the project.

5.2.2 Explanatory Power and Stability

One concern when performing regression analysis is the explanatory power of the independent variables in accounting for the variability of the dependent variable (effort). This is typically measured by R^2. However, a large value of R^2 is not the only measure of a good model. In some regards it is not even the most important. For our model $R^2 = 82.3$ and $R^2\ \text{adj} = 80.1$ which looks very good. Such a high R^2 may indicate that we have some outliers that draw the line towards them. This is closely related to model stability, which refers to the resistance to change in the fitted model under small perturbations of the data. When studying the data we observed that one project was the largest in several dimensions. Removing this project was an option we did consider. We did not remove it, but we have evaluated and

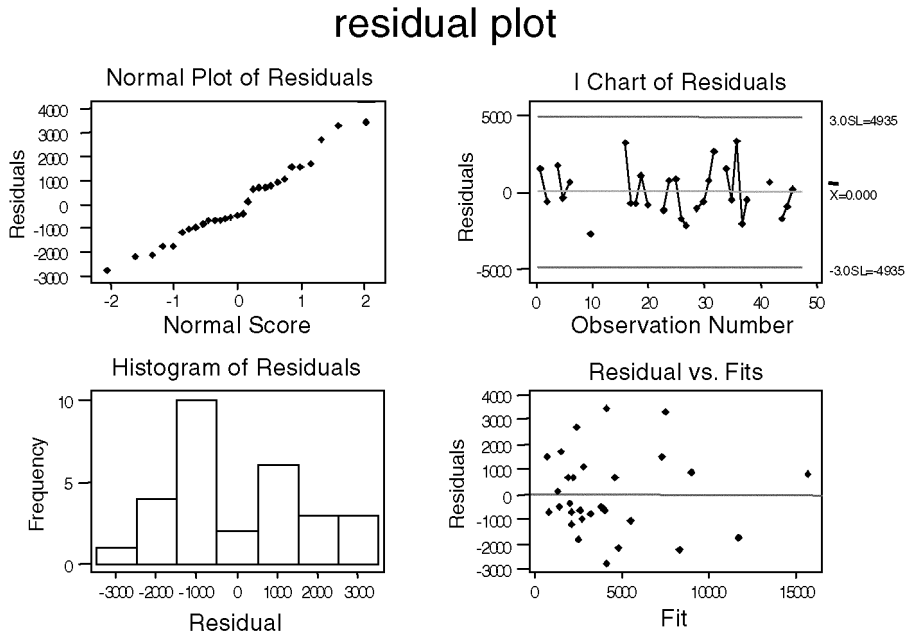

Fig. 1. Residual plots.

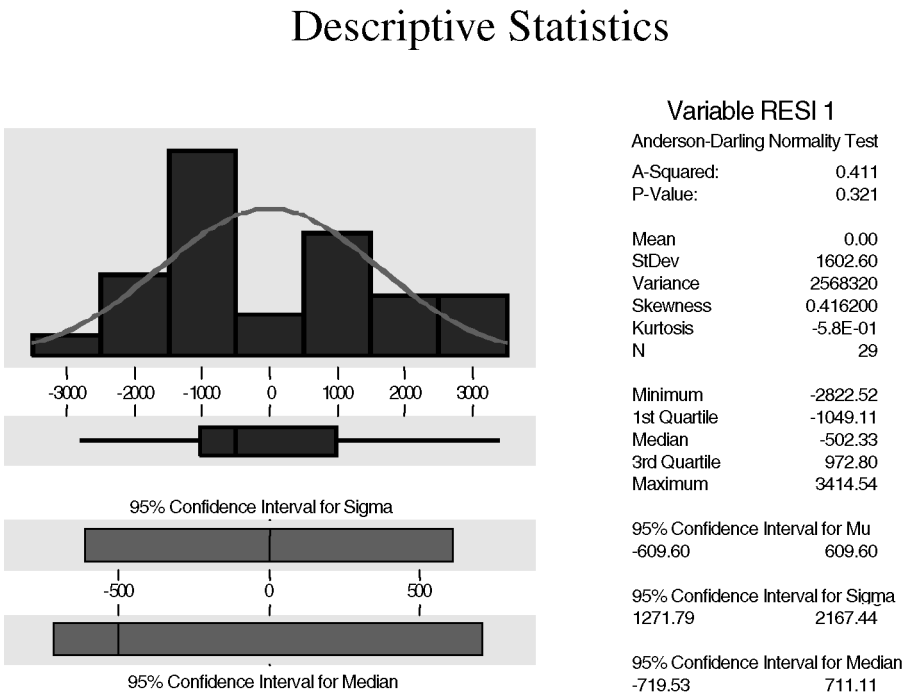

Fig. 2. Descriptive statistics for the residuals.

reported the difference it would have made as we discuss the results.

We also observed *ex post* when evaluating the results from hypothesis testing that some projects were harder to estimate than the others; both the regression model as well as the subjects made large errors. In order to compare groups of subjects, we had to remove two projects so that the two groups had equally difficult projects.

However, the model, in terms of independent variables, did not change substantially when removing any of these projects. Furthermore, R^2 adj was still high, and the model was still linear.

5.3 Assessing the Predictive Power of the Regression Model

Having established the correctness of the model, we turned our attention to assessing the predictive power of the regression model. The R^2 (adj) metric has some limitations to assess the predictive power of a regression model. Does a high R^2 (adj) mean we have an accurate model? No, the model may still produce inaccurate estimates. The reason is that R^2 (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. The true regression line (or here rather the hyperplane since we have multiple independent variables) is unknown, and the estimated value is a point estimate of the true mean. When using the model for prediction, one must consider two sources of variability. First the variability associated with the location of the true mean and secondly the variability for the probability distribution of a single value about its mean. That is, predictions far from the mean have higher uncertainty.

There are other approaches to assessing the predictive power of effort prediction models. One established measure is the Mean Magnitude of Relative Error (MMRE). The implicit assumption in this measure is that the seriousness of the absolute error is proportional to the size of the observation. For effort estimation models this seems reasonable. For example, an absolute error of four days on a small project may be comparable to an absolute error of several weeks on a much larger project.

These methods are ad hoc procedures which make no assumptions concerning the distribution of the observations. They are specifically suited to the situation at hand where the errors are increasing with the size of the observation. Indeed, if the model assumptions in the regression model have been met, the relative error measures are inappropriate since the error variance is constant. If we had transformed the variables in the model to meet the homoscedasticity assumption for regression models, the calculation of MMRE should be done using the original observations.

MMRE, therefore, provides a more realistic measure of a model's predictive power than R^2 (adj). Also, MMRE can be used to evaluate other types of models such as the analogy model whereas R^2 (adj) can only be used to evaluate regression models.

6 Experimental Design

The idea behind the experiment is the following. Let each experienced practitioner estimate a project given information about the project and the output from an analogy tool. Next, let the subject estimate a project given the same information about the project but now with the output from a regression model. Compare the estimate with the actual value and calculate the estimating error. To control for the effect of the two tools, also let the subject estimate a project using the same dataset which is available to the two tools.

The idea is that the tools add value only if the subject estimates better using the tools than when using the dataset directly. The estimating error using the dataset thus constitutes a baseline against which to assess the benefits of the two tools.

The design was influenced by a few constraints. The major risk was not getting enough subjects. We adopted two risk mitigation strategies. The first was to conduct the experiment in a classroom setting as well as via e-mail. The second was to minimize the time required by each subject. We decided that each subject should not spend more than one hour in total. We conducted a pilot study to investigate the time constraint and the overall experimental design.

6.1 Pilot Study

We first did a pilot study with seven subjects in the classroom to test some aspects of the experimental design:

- amount of time
- difficulty understanding the task
- difficulty understanding the introductions to the estimation by analogy and multiple regression models
- potential learning effects due to the sequential approach

We found that the design was reasonably realistic and did not change it. The subjects were done with each part within 10-15 minutes per part. More time did not result in better estimates for the pilot group. Also, the time did seem sufficient to understand the task and the introductions to the analogy and regression models. Last but not least, the time was short enough as to not allow the subjects to reuse knowledge from the analogy estimating part since they had to concentrate fully on understanding and using the output from the regression model in the last part. We concluded, therefore, that the design did capture the effect from the tools and not from having more time nor from learning effects. Therefore, we did not use a counterbalanced design which would have complicated the practical issues of carrying out the experiment.

6.2 Getting Subjects

We asked 118 persons to participate in the experiment. To ensure a maximum response rate we took some precautions.

To get the most senior people (and the most busy) you must be flexible. In general, it was tougher getting the most experienced personnel. Participation by e-mail ensured that we got a reasonable number of the most experienced subjects, too, since e-mail allows you to respond when and from where it suits you whereas classroom participation requires physical presence at a given time in a given location.

To get volunteers you must give something in return. We promised to give them the results, and we promised confidentiality. It is a competitive atmosphere in the organization, and everybody was keen on knowing how well they did themselves. On the other hand, nobody wants to make a fool out of himself. In addition, we gave the subjects who participated in the classroom small gifts, and we offered a free bar (after they were done!) since the experiment was done after work hours.

It helps to have high level support to promote the experiment. We ensured getting high level support and involvement from the partnership by having a few partners participate both in the classroom and via e-mail. The partnership sponsored, approved and promoted the experiment, and a few also participated themselves.

Go for the people who are interested in the outcome. We tried to get subjects who we knew were interested in project estimating. The senior personnel in particular perceive that the potential benefit is high if we improve the estimating accuracy.

Ask their opinions. People like to contribute and being asked their expert opinions. Therefore, the design of the experiment included asking their opinions in addition to gathering objective data.

6.3 Volunteer Subjects

Out of the 118 volunteers who were invited, 68 participated in the study. In a classroom experiment, 42 of them participated, while 26 volunteers participated via e-mail. The response rate via e-mail was 30 percent. All the subjects were experienced personnel with acknowledged project manager skills and at minimum 6 years of relevant practice. Many of them had 15+ years of relevant practice. All of them had previously expressed a particular interest in project estimating in some way or another. We divided the subjects into two groups based on their acknowledged experience and capability level within the organization:

- The "senior" group were subjects with 9+ years of relevant practice and 3+ years of project management and estimating practice and with one of the internal titles "Experienced Manager," "Associate Partner" or "Partner."
- The "junior" group were subjects with 6+ years of relevant practice and the title "Manager" with six exceptions who were ranked below "Manager" level.

There were 26 subjects in the "senior" group and 42 subjects in the "junior" group. The subjects represent most regions of the world: Europe, Asia, North and South America. The subjects had varying degrees of general estimating experience as well as estimating experience with this particular COTS project type (see Table 4).

6.4 Data Collection Procedure

The experimental design was motivated by a few major concerns:

- attract enough subjects with adequate seniority and experience to permit statistical analysis
- make the preparation and execution of the experiment manageable
- make a reasonably realistic experiment

We designed a synthetic environment experiment[3] to collect the data. The experiment was carried out via e-mail

3. A synthetic environment experiment is a smaller artificial setting that only approximates the environment of the larger projects [8]. The reason to use this kind of experiment is that a real environment experiment would be too expensive and time consuming.

TABLE 4
Subjects' Estimating Experience

N	Description of estimating experience
7	A: Never estimated any projects before
21	B: Estimated one or two projects before, but never any COTS project similar to this type
28	C: Estimated several projects before, but never any COTS project similar to this type
4	D: Estimated one or two COTS project similar to this type
1	E: Estimated several COTS or similar projects before, but little experience with other project types like custom.
5	F: Estimated several COTS projects and several other project types like custom or other packaged
2	Unclassified

as well as in the classroom. The e-mail subjects were offered the flexibility to answer whenever it suited them within a time frame of several months. In this way, we got more subjects than we would have got by carrying the experiment out in the classroom, only. Especially, we got more of the most experienced, and least available, practitioners.

The task the subjects were given was to provide one single number for project effort in work-days. We provided conversion rules to convert from work-hours or work-months for those used to these units. Each subject estimated the same project thrice, in three sequential parts:

- Estimating with the aid of the COTS dataset
- Estimating with the aid of an analogy tool
- Estimating with the aid of a multiple regression model

Each subject received the next part on completion of the previous part in the sequence. The projects were assigned to the subjects randomly, only ensuring that each of the projects in the dataset were assigned to at least one subject since we had more subjects than projects. Each subject estimated alone.

In Part 1 they received information about the size of the project as shown in Table 5 plus a table with 47 projects, i.e., a dataset with the project to be estimated removed from the original dataset. The table looked similar to Table 6 where we have shown a sample dataset with two projects. The dataset was provided in paper format in the classroom and electronically in bitmap format to the e-mail subjects.

In Part 2, they received the same information as in part one plus the output from the analogy tool as shown in Table 7. They did not use the analogy tool themselves. In Table 7, R1 to R10 is the ranking of the 10 closest projects, and EstDA is the estimate produced by the analogy tool.

In Part 3, they got the same information as in part one plus the output from a multiple regression model and the model itself. See Table 3. (They actually got the table in equation format which we do not show). They did not perform the regression analysis themselves.

Also, in Parts 2 and 3 the subjects received an introduction to and were explained the ideas, principles, computational algorithms as well as some strengths and limitations of the analogy tool and the regression model to let them better judge and use the output. Specifically, we pointed out that the analogy tool used unweighted dimensions.

Finally, after completing the three parts we gave them a short questionnaire to complete. The questionnaire basically asked their opinions about the perceived value of the different aids.

We gave the subjects the same information as we gave to the two tools. The only additional information given to the subjects was a list of the COTS modules. ("Modules" in Table 5 and Table 6). The two tools just got the number of modules, not the list of which modules.

We limited the time to one hour in total. This included the time for the introductions to the tools as well as for completing the questionnaire. We feared a low response rate if we had required, say, half a day or more from each subject. The classroom experiment was completed within one hour whereas the e-mail experiment ran for several months. Only, we urged the e-mail subjects not to spend more than approximately 15 minutes per part and maximum one hour in total.

6.5 Design Tradeoffs

We realize that there are a few weaknesses with the design such as potential learning effects, time and information constraints and partially a lack of control over the subjects' potential use of additional information. These weaknesses affect the realism of the experiment and thus the validity of the results. Below we discuss their potential impact on the results.

6.5.1 Learning Effects

Each subject estimated the same project both with the aid of the dataset, the analogy tool and the multiple regression tool in the same sequential order. We did not use a counterbalanced design to counteract potential learning effects, and we did not give a different project to each subject for each part.

A counterbalanced design dividing the subjects into several groups doing the parts in different sequence would have complicated the administration and execution of the experiment, especially in the classroom because we were guiding the subjects through the sequence in plenary. We wanted everybody in the same room for control reasons.

Likewise, giving the subjects the same project three times also simplified the administration since they all got a complete dataset with their own project removed. It would have been easier to make errors when distributing the datasets to the persons if we had to ensure that the dataset they got did not include any of three projects instead of only one project.

The objection to this approach is that it introduces a potential learning effect which means that performance improvements may be ascribed to the sequential order

TABLE 5
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 6
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,MM,PS,GLX**	**500**	**900**	**3400**	**4800**

TABLE 7
Output from the Analogy Tool

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**

rather than to the effect of the aids. 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.

6.5.2 Time and Information Constraints

The time and information constraints are the most severe objections regarding the realism of the experiment. To attract subjects with the desired profile we needed a flexible and not too time consuming experiment. Some objections are:

The subjects estimated alone. In practice, estimating is a group effort. At the very least, there always is a quality person checking the estimate. However, again we preferred to get a maximum of data points since the number of subjects realistically would be around 30 in each group. Below 30 data points, the statistical significance is too low.

The historical data which was given to the subjects was provided as a sheet of paper. Alternatively, we could have provided the data in a spreadsheet. This would have been OK for the e-mail subjects but more difficult for the classroom group. Also, this would easily have resulted in more time to estimate. The advantage with the data in a spreadsheet is that you can manipulate the data, e.g., by sorting on columns and rows. Therefore, the subjects would likely have got more value out of the data and probably performed better.

The subjects were given only the outputs from the analogy tool and the multiple regression analysis. They were not given the tools themselves. Using the tools would presumably have improved their understanding and thus their performance with the aid of the tools. Again, the time constraints did not permit this approach.

The introduction to the principles and workings of the estimation by analogy and the regression analysis was brief, approximately 20 minutes for the classroom subjects. A more thorough presentation would probably have resulted in better performance using the tools. Especially this may be true for the analogy tool which was unfamiliar to everybody unlike regression analysis which several of the subjects were familiar with.

The subjects received the same information as the tools with one exception since we wanted to compare human and tool performance based on some idea of "fairness." However, in practice practitioners use additional information, e.g., in prose format, that a tool cannot use.

6.5.3 Lack of Control of Information Use

We did not have any control over the e-mail subjects. They could easily spend more time or access additional information as opposed to the classroom subjects were we were present and could control cheating. We believe, however, that this is not a big problem since most of these subjects were very senior persons with limited time. Also, by comparing the performance of the e-mail and classroom groups, we could have discovered if the e-mail subjects had cheated.

6.5.4 Summary of Design Discussion

There has been a difficult tradeoff between on one side getting enough subjects and making the experiment manageable and on the other side making it realistic. We believe that the design is reasonable for relative comparisons of the three aids dataset, analogy tool, and regression model. However, we believe practitioners would perform better in a real, rather than in this synthetic, environment. The design will be further discussed in Section 11, Discussion.

Finally, the estimates were *ex post*. In real life, estimates are done *ex ante*. When you know your budget it is more likely that the actual value is closer to the estimate than if not since the job of the project manager is to manage to the budget. Therefore, an *ex ante* estimating experiment would yield higher accuracy levels. The downside is that this would require a multiyear longitudinal experiment. For the purpose of evaluating the two tools against using history and against each other, this design should be acceptable.

7 Test Metrics

We used the Magnitude of Relative Error (MRE) as the fundamental test metric. MRE is defined the usual way:

$$\mathrm{MRE} = 100 \times \left| \frac{(\text{Actual Effort} - \text{Estimated Effort})}{\text{Actual Effort}} \right|$$

MRE is a general as well as a reasonable test metric. In general, 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. Furthermore, measuring the relative error seems to be more appropriate than measuring the absolute error since a budgeted 10 million dollar project in most cases can cope with a 1 million dollar overrun better than a 1 million dollar project can. Furthermore, buyers usually accept accuracy figures in percent understanding that a large project cannot be estimated with the same absolute accuracy as a small project.

The estimating accuracy was evaluated with two test metrics:

- Mean Magnitude of Relative Error (MMRE) and
- Median Magnitude of Relative Error (MdMRE).

The major advantage of the mean over the median as a measure of central tendency is that it takes into account the numerical value of every single observation in the data distribution. It represents a balance point, or center of gravity, in that the sum of the distances to the observations below it, is equal to the sum of the distances to the observations above it. This mathematical characteristic of the mean makes it a cornerstone of statistical analysis. Ironically, its sensitivity to every numerical value becomes its chief drawback in skewed distributions and when there are one or two outliers among the observations.

We use the median in addition to the mean. Aside from its common sense interpretation as a truly central value, the most attractive feature of the median is its insensitivity to the values of the very extreme scores in a distribution, which are atypical and sometimes flukes. The median is thus especially appropriate as a measure of a central tendency of a skewed distribution of data. Its main limitation is that it does not have mathematical properties that lend itself to more advanced analyses. This limitation does not, however, affect our relatively simple analyses.

The statistical significance of the results were tested using:

- a T-test of mean difference between paired MREs (for MMRE)
- a Wilcoxon Rank Sum Test or Mann-Whitney U Test (for MdMRE)

The significance tests are vital since we cannot conclude solely by observing differences in means or medians. These differences could have been caused by chance alone because samples drawn from the same underlying population might have different means or medians.

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 8 we see that $\mathrm{MMRE_A} = 154$ percent and $\mathrm{MMRE_R} = 127$ percent. So apparently, the multiple regression tool outperforms analogy tool for this dataset. 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.

Wilcoxon Rank Sum Test (or Mann Whitney Test). Another way of testing the results is by performing a Wilcoxon test. Wilcoxon performs a one-sample rank test of the median. The test assumes that the data are a random sample from a symmetric population. This test is slightly less powerful than the t-test if the population is normal, while it may be considerably more powerful for other populations.

In addition to mean and median accuracy figures, practitioners are concerned with the risk of making very erroneous estimates that deviate substantially from the expected inaccuracy. This was evaluated with two additional test metrics:

- The Standard Deviation of $\mathrm{MRE}(\mathrm{SD_{MRE}})$
- The Maximum Magnitude of Relative Error $(\mathrm{MAX_{MRE}})$

8 Initial Results

In this and the following section, we present only those results that are sensitive to the experimental design , i.e., the type and size of the dataset, the norms for removing outliers and other data points from the original dataset, the test metrics, significance levels, and least but not last, the use of human subjects and their level of expertise. Thus, the main motivation for presenting the results is to enable us to discuss the problems of convergence and to explain disagreements with previous studies [4], [5], [6], [7]. A more detailed presentation of the results is outside the scope of this paper and may be found in [6], [7].

Table 8, Table 9, and Table 10 show human and tool performances. Parts 1, 2, and 3 show *human* performance when estimating with the aid of a history (Part 1), an analogy tool (Part 2) and a multiple regression tool (Part 3), respectively. The analogy and multiple regression *tool* performances are shown in the rows ANGEL and MR, respectively.

We have divided the subjects into two groups based on their experience level. Table 8 shows results for both groups together whereas Table 9 and Table 10 show results for each group. The first group with approximately 6-9 years experience is named "junior." The other group, i.e., the

TABLE 8
Estimating Performance Results—All Subjects

	N	MMRE %	MdMRE %	SD_{MRE} %	MAX_{MRE} %
part 1 all	**61**	**243**	**59**	**628**	**3900**
Part 2 all	**58**	**136**	**51**	**230**	**1208**
Part 3-all	**56**	**126**	**43**	**192**	**900**
ANGEL all	**68**	**154**	**52**	**303**	**1476**
MR all	**68**	**127**	**35**	**227**	**1051**

N *is number of observations*

TABLE 9
Estimating Performance Results—Junior Group

	N	MMRE %	MdMRE %	SD_{MRE} %	MAX_{MRE} %
Part1 jun	**40**	**321**	**61**	**762**	**3900**
Part2 jun	**39**	**173**	**59**	**269**	**1208**
Part 3 jun	**38**	**154**	**50**	**217**	**900**
ANGEL jun	**41**	**177**	**55**	**310**	**1476**
MR- jun	**41**	**169**	**45**	**274**	**1051**

N *is number of observations*

TABLE 10
Estimating Performance Results—Senior Group

	N	MMRE %	MdMRE %	SD_{MRE} %	MAX_{MRE} %
Part1 sen	**21**	**94**	**52**	**113**	**355**
Part2 - sen	**19**	**60**	**43**	**78**	**350**
Part 3 sen	**18**	**67**	**28**	**107**	**414**
ANGEL-sen	**27**	**119**	**49**	**294**	**1476**
MR-sen	**27**	**64**	**27**	**104**	**478**

N *is number of observations*

most experienced people with 9+ years of experience, is named "senior."

The most important observation in Table 8 from a method perspective is that some of the preliminary results depend on whether we use MMRE or MdMRE as test metric. Dramatic differences in terms of MMRE are less dramatic in terms of MdMRE. Comparing the performance of practitioners using the regression model and the performance of the regression model itself, we find that practitioners using the regression model perform better than the regression model itself using MMRE as test metric (126 percent vs. 127 percent) whereas the result is the opposite using MdMRE as test metric (35 percent vs. 43 percent). The other results in Table 8 are consistent across both the test metrics MMRE and MdMRE. This is true also for the SD and MAX figures.

Table 9 shows a similar overall tendency as Table 8. An interesting observation in Table 9 is that when comparing the performance of the subjects using the analogy tool with their performance using the dataset alone, the MMRE results suggest that they improve by a factor of two (173 percent vs. 321 percent). Using the MdMRE the results are much less impressive (59 percent vs. 61 percent). This indicates that the MMRE results may be biased by a few extreme observations, and this is indeed confirmed by the MAX_{MRE} results showing that the extreme errors are reduced by a factor of three when using the analogy tool compared with using the dataset alone.

Also, using MMRE we observe again that junior practitioners using the analogy tool perform better than the analogy tool itself (173 percent vs. 177 percent) but that this result is contradicted using MdMRE (59 percent vs. 55 percent). The same observation of inconsistency applies when comparing human and tool performance for the regression model.

Similarly, the MMRE results in Table 10 suggest that senior practitioners benefit more from the analogy tool than from the regression model (60 percent vs. 67 percent). This result is contradicted by the MdMRE results (43 percent vs. 28 percent). Also, seniors using the analogy tool seem to perform twice as good as the analogy tool itself in terms of MMRE (60 percent vs. 119 percent). The MdMRE results confirm this result to a lesser extent (43 percent vs. 49 percent).

Comparing juniors with seniors the results suggest that seniors estimate more accurately and more consistently than juniors. Using MMRE the seniors really seem to outperform the juniors (e.g., 94 percent vs. 321 percent using the dataset). The results are less impressive in favor of the seniors when using MdMRE (52 percent vs. 61 percent). This suggests a few extreme observations in the junior group. This is indeed supported by the SD and MAX results (e.g., 355 percent vs. 3,900 percent using the dataset).

TABLE 11
Significance of Results—T-Test of Mean of Paired Differences

	H1	H2	H3	H4	H5	H6
	$MMRE_2 > MMRE_3$	$MMRE_2 > MMRE_4$	$MMRE_4 > MMRE_3$	$MMRE_A < MMRE_3$	$MMRE_R < MMRE_4$	$MMRE_R > MMRE_A$
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 asterisk $(*)$ and $(**)$ mean statistically signficant at $\alpha = 10\ percent$, respectively.

TABLE 12
Significance of Results—Wilcoxon Rank Test of Median

	H1	H2	H3	H4	H5	H6
	$MdMRE_2 > MdMRE_3$	$MdMRE_2 > MdMRE_4$	$MdMRE_4 < MdMRE_3$	$MdMRE_A < MdMRE_3$	$MdMRE_R < MdMRE_4$	$MdMRE_R < MdMRE_A$
All	**6 (0.04)****	**6 (0.05)****	**1 (0.29)**	**2.5 (0.26)**	**4.5 (0.2)**	**9.5 (0.05)****
Senior group	**7 (0.08)***	**7 (0.06)***	**1 (0.21)**	**6.5 (0.19)**	**4.5 (0.27)**	**6 (0.3)**
Junior group	**5 (0.17)**	**4 (0.3)**	**0 (0.5)**	**0 (0.62)**	**4 (0.31)**	**17.5 (0.01)****

The number not in parenthesis is the median, and the number in parenthesis is the significance level. The astericks $(*)$ and $(**)$ mean statistically significant at $\alpha = 10\ percent$ and $\alpha = 5\ percent$, respectively.

However, the most important observation is that the seniors got easier projects to estimate despite the random assignment of projects to subjects. Using the preliminary results (i.e., before significance testing) from the regression tool itself as the norm and using MMRE as test metric, it seems that the juniors got twice as difficult projects as the seniors (169 percent vs. 64 percent in the MR rows of Table 9 and Table 10). Therefore, the comparison of junior and senior performance is invalid.

However strong the results might appear in terms of MMRE and MdMRE the results above are only preliminary. We cannot conclude without testing the significance of the results because samples with different means or medians may still have been drawn from the same population.

Using MMRE as test metric and a t-test of the mean of paired differences we find that most of the results are not significant at a 10 percent, and none are significant at a 5 percent level. This is shown in Table 11 which shows the mean of paired differences with the significance level in parenthesis.

One important observation is that apparently strong results in terms of MMRE differences are not significant. For example, comparing the performance of seniors using the regression model with their performance using the dataset they seem to benefit substantially from the regression model in terms of MMRE (67 percent vs. 94 percent). However, Table 11 tells that this result is not significant (H2 column, Senior group row). Also, seemingly very strong results such as comparing junior performance using regression with using the dataset (MMRE: 154 percent vs. 321 percent in Table 9) are significant at a 10 percent level, only, but not at a 5 percent level.

Another way of testing the results is by performing a Wilcoxon test. See Table 12. The most important observation comparing Table 11 and Table 12 is that the t-test in Table 11 suggests that juniors benefit *significantly* using the analogy tool compared with using the dataset. Using the Wilcoxon test in Table 12 this result is no longer significant.

On the other hand, the t-test does not suggest that seniors benefit significantly from the regression model whereas the Wilcoxon test says they do. The Wilcoxon test suggests stronger that all subjects benefit from both tools compared with using the dataset where the Wilcoxon test is significant at the 5 percent level whereas the t-test is significant at the 10 percent level, only.

Another interesting result is that the regression model does not outperform the analogy tool using the t-test whereas it does outperform the analogy tool at the 5 percent level using the Wilcoxon test.

9 Adjusted Results

The assignment of projects to subjects was random. Specifically, the assignment of projects to subjects in the senior and junior groups was random. *Ex ante* we did not know who were seniors and juniors. (This was the first question they answered as they took part in the experiment.) *Ex post*, however, using the regression model as a norm we observed that the seniors seemed to have been assigned easier projects (See Initial Results section). The two most difficult projects were assigned to juniors. MMRE for the regression model are 64 percent for the seniors and 169 percent for the juniors which is a significant difference. When removing the two most difficult projects, the groups got comparable projects. The difference using the regression

TABLE 13
Adjusted Estimating Performance Results—Junior Group

	N	MMRE %	SD_{MRE} %	MAX_{MRE} %
Part1 jun	**38**	**257**	**645**	**3900**
Part2 jun	**37**	**142**	**216**	**900**
Part 3 jun	**36**	**142**	**216**	**900**
ANGEL jun	**38**	**144**	**280**	**1476**
MR- jun	**38**	**108**	**172**	**783**

N *is number of observations*

model as a norm is now 64 percent for seniors and 108 percent for juniors which is no longer a significant difference.

The overall trend for the juniors in Table 13 is similar to the trend in Table 9 except that in Table 13 the performance is equal using both analogy and regression tools (142 percent) whereas in Table 9 the juniors seem to perform better using the regression model than when using the analogy tool.

Also, the results in Table 13 suggest that the regression tool itself seems to outperform the analogy tool more than it does in Table 9. However, this is not surprising since we have used the regression model as the norm for identifying and removing " difficult " projects.

The most important observation when comparing the t-tests in Table 14 with Table 11 is that results that were not significant in Table 11 now have become significant in Table 14. Removing two projects from the sample totally alters the results. The regression tool estimates better than juniors aided by the same tool (H5), and this result is now significant at the 5 percent level whereas it was not significant before. The result is similar when comparing regression and analogy tool performance (H6) for the whole group of subjects (All). However, again we have to be cautious with the conclusions since we have used the regression model as the norm for removing observations from the sample.

Similar observations apply when comparing the results of the Wilcoxon tests in Table 12 and Table 15. In Table 12 the juniors did not perform significantly better using the analogy tool. In Table 15 they do perform better at the 10 percent level. Using the regression model the juniors now perform significantly better at the 5 percent level compared with not significant in the initial results.

More surprising, the Wilcoxon test suggests that the regression tool now does not outperform the analogy tool whereas in the initial results the regression tool outperformed the analogy tool at the 5 percent level for the junior observations. This is surprising since we used the regression model as the norm for "improving" the sample of observations.

In summary, the only really robust results across perturbations of the dataset and across the two test metrics is that both the analogy tool and the regression model improve human performance compared with human performance using the dataset alone. That is, H1 and H2 are confirmed.

Another robust result is that the analogy tool does not outperform the regression model (H6). Rather, the results suggest an opposite trend. However, this result is based on using the regression model as the norm when selecting the dataset on which to validate the two tools.

10 Results of Questionnaire

This section is included partly as a curiosity and partly to point out that bridging the gap between research and practice requires marketing effort in addition to good

TABLE 14
Adjusted Significance of Results—T-Test of Mean of Paired Differences

	H1	H2	H3	H4	H5	H6
	$MMRE_2 > MMRE_3$	$MMRE_2 > MMRE_4$	$MMRE_4 < MMRE_3$	$MMRE_A < MMRE_3$	$MMRE_R < MMRE_4$	$MMRE_R < MMRE_A$
All	**72 (0.1)***	**69 (0.11)**	**-1.8 (0.58)**	**3.4 (0.45)**	**29 (0.04)****	**44 (0.07)***
Senior group	**20.7 (0.06)***	**1.7 (0.44)**	**-10.8 (0.18)**	**-12.2 (0.76)**	**2.28 (0.82)**	**5.5 (0.11)**
Junior group	**98 (0.12)**	**103 (0.12)**	**3 (0.41)**	**11 (0.4)**	**42 (0.04)****	**35 (0.19)**

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

TABLE 15
Adjusted Significance of Results—Wilcoxon Rank Test of Median

	H1	H2	H3	H4	H5	H6
	$MdMRE_2 > MdMRE_3$	$MdMRE_2 > MdMRE_4$	$MdMRE_4 < MdMRE_3$	$MdMRE_A < MdMRE_3$	$MdMRE_R < MdMRE_4$	$MdMRE_R < MdMRE_A$
Senior group	**7 (0.08)***	**7 (0.06)***	**1 (0.21)**	**6.5 (0.19)**	**4.5 (0.27)**	**6 (0.3)**
Junior group	**6 (0.07)***	**7 (0.05)****	**1 (0.2)**	**6.5 (0.18)**	**13 (0.1)***	**11 (0.15)**

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

TABLE 16
Practitiones, Tool Preferences

	Prefer ANGEL	Prefer MR	History
all	**11**	**13**	**17**

products. The whole experiment was partly intended as a marketing effort.

Table 16 shows that the subjects prefer history to both tools. Table 17 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 indicate that the tools actually add value.2

We find it interesting that the objective estimating results differ from the subjects' 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.

11 Discussion

In this section, we discuss the problems of convergence and explain disagreements with previous studies [4], [5], [6], [7]. The preceding Results sections suggest that the results are sensitive to a number of factors, in particular to the data, the experimental setup and the data analysis. The experimental setup extends previous studies by measuring the performance of the *toolusers* in addition to measuring the performance of tools in isolation from the users. There are thus several reasons why perfect replication with [4], [5] was not achieved.

11.1 Data

The norms for cleaning the data, i.e., for removing outliers and data points with missing values, affects the results as to "which tool is best." We used the limitations of the regression model as the norm for removing outliers and data points with missing values. This norm favors the regression model. We do not know which data cleaning procedure and norms Shepperd et al. used. Also, the original dataset contains more information in prose format that probably would add value to a human estimator but which regression and analogy tools cannot use. We gave the human subjects the same information as we gave to the tools with one exception. The human subjects received information about which COTS modules were in scope in addition to the counts of modules (see Table 5 and Table 6). We also found that our own results did not converge for two slightly different datasets. Removing two data points more from the initial dataset completely altered the results (compare Table 11 and Table 14).

The number of data points will likely impact on the results regarding "how good is the tool." Our dataset contained 48 data points. A dataset with 10 data points is easier to inspect visually by a human subject than a dataset with a thousand projects. We perceive the value of ANGEL to be mainly in ranking the projects with respect to closeness with the project to be estimated. With less than 10 projects in the sample, we believe that ANGEL would not add significant value to a human estimator. As for regression, it does not make sense to produce a regression line with very few data points.

The number of independent variables will likely impact on the results regarding "how good is the tool." Our dataset is multidimensional with 10 independent variables. A dataset with one independent variable such as function points and one dependent variable such as effort would be easy to inspect visually by a human subject using a spreadsheet with simple sort functionality. It seems obvious that the added value of regression and analogy tools will be higher for multidimensional datasets than for two-dimensional datasets.

The interval between the smallest and largest projects will probably impact on "which tool is best." Our dataset spans industrial COTS projects from 100 to 20,000 workdays. Regression models extrapolate from the data points. Thus, a dataset that spans a large size interval such as ours probably is best analyzed with regression whereas a dataset where the data points are lumped together probably favors the analogy approach. Consider the extreme case where all data points were, say, 4,000 workdays. In this case, analogy would produce an $\mathrm{MMRE} = 0$ percent. On the other hand, regression would not be able to make any sense at all out of such a dataset. We suspect that the datasets used by Shepperd et al. are lumped more together than ours.

The dataset homogeneity will impact on the absolute performance, i.e., "how good is the tool." Our dataset is homogeneous in that it includes projects where the COTS package is from one single vendor, only. However, we included both new development projects and later release projects to get enough data points in the sample. We do not know anything about the degree of homogeneity of datasets used in Shepperd et al.'s studies.

The general data quality will impact on results with respect to "which tool is best" and "how good is the tool." Our COTS dataset was collected by people in one organization using a standard methodology. This suggests a reasonably high interrater reliability of size counts and consistent effort measurements. We experienced that it is a tremendous task

TABLE 17
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**

to gather reliable data even within one single homogeneous organization, and we suspect that lack of data quality is one of the major flaws of all empirical science. If your data are rubbish, so are your results with respect to "which tool is best" as well as "how good is the tool."

Furthermore, *how representative of the population is your sample* of data. Our COTS dataset consisted of mainly United States projects. Project team characteristics and client characteristics may result in different productivity in, say, the United States and Norway. This means there is a risk that estimating a Norwegian project will be less accurate than the expected inaccuracy we have found (MMRE or MdMRE). Finally, few datasets fully satisfy all the technical requirements of data analysis methods, e.g., complying with normal distribution requirements. This means your hard numbers are not that hard, after all.

11.2 Experimental Setup

Our main contribution to improving experiments to validate estimating models and tools is that we used human subjects, experienced practitioners, in the experiment. To our knowledge, no previous studies have validated estimating models this way.

Extending previous experiments by using human subjects impacts on the results as to "how good is the tool" and "which tool is best." As to "which tool is best," we found that the regression model outperforms the analogy tool when testing tool performance alone and using MdMRE as the test metric and the Wilcoxon rank test. (see Table 12, column H6, row "All"). This result is significant at the 5 percent level. Testing human performance, neither tool was a superior aid when using MdMRE as test metric (see Table 12, column H3, row "All"). This result suggests, therefore, that the best tool when tested alone is not necessarily the best tool for a human subject. Testing tool performance alone would be justified only if it turns out that tools always outperform people using the tools, so that tools could replace people. As for "how good is the tool," we found that the results depend on whether human subjects are involved or not. For example, $\mathrm{MMRE} = 60\ \mathrm{percent}$ for seniors using ANGEL whereas $\mathrm{MMRE} = 119\ \mathrm{percent}$ for ANGEL alone (see Table 10).

When using human subjects their skill level impacts on the results as to " which tool is best. " We found that juniors benefit from the regression model whereas seniors do not. (Table 11, column H2, rows "senior" and "junior").

When the tool validations involve human subjects, there arises a number of design trade-offs regarding the experimental setup as discussed in Section 6.5 Design Trade-Offs. In addition, in a real environment (as opposed to a synthetic environment) estimating is performed *ex ante*, not *ex post*, and part of the project manager's responsibility is not only to estimate accurately but also manage the project to the budget or estimate. Therefore, the results as to "how good is the tool" likely will be (very) different in the two cases. This implies that estimating performance cannot be seen in isolation from project management performance. Thus, research effort aimed at improving estimating accuracy must also address improving the planning, control, and execution of projects.

The results are very conservative with respect to human performance. The design in general was biased towards giving the tools optimal conditions. In practice, estimating is a more time consuming activity. For example, 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.

11.3 Data Analysis

The results do not converge across test metrics, significance levels and after the removal of outlier results.

The results do not converge for the various test metrics (MMRE, MdMRE, SD, and MAX). In particular, they do not converge for the mean and the median MRE which both are reasonable measures of a central tendency. We used partially different test metrics from Shepperd et al. They used MMRE and PRED. The results are also sensitive to the significance levels chosen and to using vs. not using significance tests. We used two different significance tests, t-test for the mean and Wilcoxon test for the median, and we reported the results for the 5 percent and 10 percent significance levels. Shepperd et al. did not report significance tests.

When estimating using ANGEL, it finds the best subset independent variables by tuning it to either MMRE or PRED. We used MMRE. For the regression model the best subset of independent variables was found using R^2. This favors ANGEL when MMRE is used as test metric. However, when using the MdMRE none of the tools are favored.

One unexpected problem we ran into was that the random assignment of projects to subjects did not result in giving equally difficult projects to the two groups as we expected. We observed this by using the regression model and MMRE as the norm. (See Table 9 and Table 10, column MMRE, row MR, where MMRE is 169 percent vs. 64 percent). When removing the two "outliers" the results are altered. The regression tool now performs significantly better than ANGEL (at a 10 percent significance level, see Table 14). On the other hand, using ANGEL as the norm and the other test metrics (MdMRE, $\mathrm{SD}_{\mathrm{MRE}}$, and $\mathrm{MAX}_{\mathrm{MRE}}$) the two groups seem to have received equally difficult projects. In this case, the final results would have been identical to the initial results. This means that the adjusted results as to "which tool is best" favors the regression model when using the MMRE. This also shows how sensitive the results are to the data analysis.

12 Conclusion

To have general validity, empirical results must converge. Shepperd et al. [4], [5] claim that analogy outperforms regression. We have shown that this claim does not have general validity. However, our results do not imply that estimation by analogy may not be a superior model to regression models in other settings than ours. Furthermore, our results suggest that both models, analogy and regression, actually add value to industrial, experienced practitioners if the setting is similar to ours.

A more important observation than the results themselves is that we have shown that the results are sensitive to the experimental design: the data, the experimental setup and the data analysis. To be credible, an experimental science must understand the limitations of models and theories and the limitations of empirical methods and techniques. Understanding the limitations of theories and empirical methods is key to be able to explain the disagreements of empirical results. In empirical science there are many pitfalls that may impact on the results.

We have shown that the results are sensitive to the data such as the norms for cleaning the data, the number of data points, the number of independent variables, the interval between the smallest and largest projects, the dataset homogeneity and the general data quality. Based on the results of this research as well as our project management and estimating experience our gut feeling is that major improvements in estimating accuracy must come from improving the quality of the historical data, in particular the consistency of effort data across projects and organizations. We suspect that poor data quality is one of the major flaws of all empirical science.

We have shown that the results are sensitive to the experimental setup. Our main contribution to improving experiments to validate estimating models and tools is that we used human subjects, experienced practitioners, in the experiment. To our knowledge, no previous studies have validated estimating models this way. However, we acknowledge that validating estimating models by testing tool performance in isolation from human subjects is a useful, and definitely cheaper, first screening mechanism. Nevertheless, we argue that the ultimate test of an estimating tool must include representative users. Our results suggest that the answers to "which tool is best" and "how good is the tool" depends on the question "for whom." To put it in the words of Brooks: "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]. "

We have shown that the results are sensitive to the data analysis such as the test metrics and the significance tests and levels. It is easy to make errors in all steps of an empirical endeavor. Ironically, modern statistical analysis tools like SPSS, SAS, and MiniTab make it easy, too easy, to produce a regression model. It is easy to do, and to impress with, all the technicalities. The hard part is to gather and select the right data and to select the appropriate analysis methods and test metrics. Probably, the main lesson learned is this: be critical, very critical, to empirical results. Also, be prepared for work, much work, in carrying out empirical research. Don't forget your basic assumptions and don't generalize your results beyond their limitations.

Finally, we should not forget the ultimate purpose of this research when diving into the meticulous empirical task. As practitioners, we are continuously searching for aids that will improve our performance. In this study, the ultimate question we are seeking an answer to is a simple one: "Should I use ANGEL (or multiple regression) or should I not? Does it add value?" Unfortunately, statistical methods rarely provide you with binary "true" or "false" answers. All they provide you with are likelihoods and a few figures such as a couple of MMREs and a "p value." Therefore, this information is just one of several inputs we need to make a decision. Other aspects impacting on the final decision include gut feeling and knowledge of the data, the user and the project you are to estimate.

ACKNOWLEDGMENTS

The authors are most grateful to all the subjects, our former colleagues at Andersen Consulting, who volunteered in the experiment. We would also like to thank the numerous knowledge champions at Andersen Consulting's global SAP service line who helped us improve the quality of the COTS dataset and the partnership in Andersen Consulting who sponsored the research. Last but not least, we would like to thank the guest editors and anonymous referees. Their suggestions contributed to improving the focus and clarity of the paper.

REFERENCES

[1] *COCOMO II Model Definition Manual,* version 1.4, Univ. of Southern California, http://sunset.usc.edu/COCOMOII/cocomo.html, 1997.

[2] F.P. Brooks, "The Computer Scientist as Toolsmith II," *Comm. ACM,* vol. 39, no. 3, pp. 61–68, Mar. 1996.

[3] *IFPUG Function Point Counting Practices: Manual Release 4.0,* Int'l Function Point Users' Group, Westerville, Ohio, 1994.

[4] M.J. Shepperd, C. Schofield, and B.A. Kitchenham, "Effort Estimation Using Analogy," *Proc. 18th Int'l Conf. Software Eng.,* pp. 170–178, Berlin, Mar. 1996.

[5] M.J. Shepperd and C. Schofield, "Estimating Software Project Effort Using Analogies," *IEEE Trans. Software Eng.,* vol. 23, no. 12, pp. 736–743, Nov. 1997.

[6] E. Stensrud and I. Myrtveit, "The Added Value of Estimation by Analogy—An Industrial Experiment, " *Proc. The European Software Measurement Conf., FESMA'98,* pp. 549–556, Antwerp, Belgium, May 1998.

[7] E. Stensrud and I. Myrtveit, "Human Performance Estimating with Analogy and Regression Models: An Empirical Validation," *Proc. Fifth Int'l Symp. Software Metrics, METRICS'98,* pp. 205–213, Bethesda, Md., Nov. 1998.

[8] M.V. Zelkowitz and D.R. Wallace, "Experimental Models for Validating Technology," *Computer,* vol. 31, no. 5, pp. 23–31, May 1998.

Ingunn Myrtveit received her MS degree in management from the Norwegian School of Management in 1985 and her PhD degree in economics from the Norwegian School of Economics and Business Administration in 1995. Dr. Myrtveit is an associate professor in business economics at the Norwegian School of Management. She conducted research for this article while a senior manager at Andersen Consulting's World Headquarters R&D Center in Chicago. Her research interests include managerial economics, incentives, and empirical studies.

Erik Stensrud received an MS degree in physics from The Norwegian Institute of Technology in 1982 and an MS degree in petroleum economics from the Institut Francais du Petrole in 1984. Stensrud is currently a senior manager at Ernst & Young Consulting. He conducted part of the research for this article while a senior manager at Andersen Consulting's World Headquarters R&D Center in Chicago. His research interests are in software engineering and project management and the application of empirical studies and software metrics within these areas. Stensrud is a member of the IEEE, IEEE Computer Society, ACM, and The Norwegian Computer Society.