

To Estimate or #NoEstimates, that is the Question

Abstract

A common approach in agile projects is to use story points, velocity and burnup charts to provide a means for predicting release date or project scope. Another approach that is proposed is to abandon story point estimation and just count stories using a similar burnup chart. We analyzed project data from 55 projects claiming to use agile methods to investigate the predictive value of story point estimation and velocity for project forecasts. The data came from nine organizations ranging from startups to large multinational enterprises. We found that projections based on throughput (story counts) were essentially identical to that of using velocity (story points). Neither velocity nor throughput were great predictors as the uncertainty bands were rather large. Through the use of a simulation model we replicated our findings which aid in understanding the boundary conditions for when story point estimates may be better predictors.

1. Introduction

TO Estimate, or #NoEstimates, that is the question:
Whether 'tis nobler in the mind to suffer

The Slings and Arrows of outrageous errors,
Or to take Arms against a Sea of troubles,
And by opposing, end them? #NoEstimates, so easy;
No more; and by #NoEstimates, focus on value and end

The headaches and the thousands of natural unknowns
That estimation is prone to, 'tis a consumption
Devouring all resources. #NoEstimates; so easy;
So easy; perchance it seems: ay, there's the rub;
For in #NoEstimates what surprises may come,
When we develop in earnest and begin our toil.
Must give us pause. There's the respect
That makes calamity of software's life;

Since the early days of software development, teams have struggled with estimation challenges. Many have searched for the holy grail of processes or tools to make us better estimators, to limited avail. Recently, the #NoEstimates movement in the twitterverse has challenged some of these fundamental tenets. As defined by originator Woody Zuill:

“#NoEstimates is a hashtag for the topic of exploring alternatives to estimates [of time, effort, cost] for making decisions in software development. That is, ways to make decisions with ‘No Estimates’.”[1]

A standard approach for agile/Scrum teams is to use story points and velocity [2] to track progress. Story points are an estimate of the relative effort required for a given user story and are only meaningful within a particular team. A team is free to use whatever means they want to obtain relative sizes of story points. A team could chose to use something formal like Function Points [3], but this is rather rare. Most teams will use a relative effort scale based on the team judgement. Velocity is defined as story points completed per time unit or iteration. Just as with a car we have instantaneous velocity of the iteration, and the average velocity across multiple iterations. Using the average velocity the team can use a burnup or burndown chart to extrapolate the time required to complete the remaining stories in the backlog, or to determine the approximate number of story points that can be completed in a fixed time [4]. The burnup chart in Figure 1 shows this extrapolation. The use of a burnup or burndown chart is mathematically identical to more traditional project management tools such as Earned Value Management (EVM) [5]. One approach often suggested by #NoEstimates advocates is to discontinue estimating story points and instead simply count the number of stories completed per iteration, also called throughput [6].

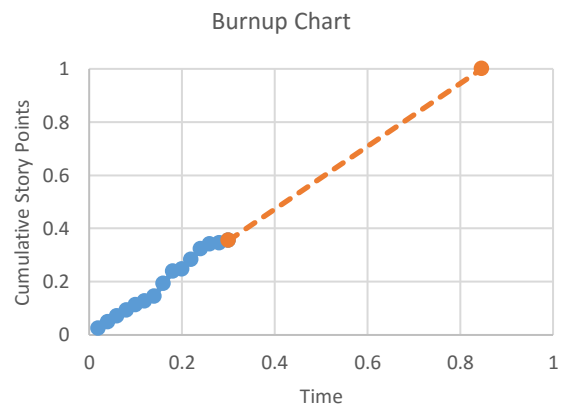


Figure 1

Many of the claims by #NoEstimates advocates are anecdotal. We wanted to see real data. So to examine the predictive value of velocity and throughput, we analyzed data [7] from 55 projects claiming to use agile methods. The data came from 9 different organizations ranging from startups to large multinationals.

In order to explore the parameters that influence the predictive value, we built a simulation model of the agile estimation and delivery process. We first validated that we could replicate the findings of the data and then explored which parameters might impact the usefulness of the estimates.

For many of our comparisons we make use of the P90/P10 ratio. This ratio is commonly used in domains where the range of the distribution is rather large. The P90 is the 90th percentile and P10 is the 10th percentile. Some domains where it is used frequently are wealth distributions and oil and gas exploration. As a concrete example, the P90/P10 ratio for income inequality is, roughly speaking, the ratio of what professionals like doctors and lawyers earn to what cleaners and fast food workers earn [8]. One of the nice attributes of the P90/P10 ratio is that it uniquely describes the distribution shape for either a lognormal or Weibull distribution.

2. Core findings

Our analysis of the data showed that for the purpose of tracking progress and projecting into the future, there was no significant advantage to using story point estimates and velocity over tracking throughput. We also replicated this in simulations.

By altering the simulation conditions, we explored when story point velocity is a better predictor than throughput. The key parameters we altered were:

- Distribution of story size
- The team's estimation accuracy
- The bucketing approach used (e.g none, modified Fibonacci, power of 2, etc)

It turned out that story point estimation adds some value when there is large variation in story size. Bucketing had very little impact although larger buckets such as powers of 4 slightly eroded the added value of velocity.

2.1 Hardening

Neither velocity nor throughput correctly accounted for what we believe to be a hardening period. Approximately 50% of the projects reported a 2-12% timespan at the end of the project with no stories delivered. In five projects we observed zero velocity from 30%-50% of the overall schedule. It did not seem right to use them in the analysis as agile development encourages continuous delivery. Thus they were excluded from further analysis.

2.2 Velocity distribution and estimation accuracy

The variability across all the projects and all the iterations of the instantaneous velocity is a direct measure of estimation accuracy of the story points. The data had a P90/P10 ratio of 4.5 and exhibited characteristics of a lognormal or Weibull (failure) distribution. This large of a distribution ratio indicates that velocity is not very consistent from iteration to iteration. We also discovered that this range did not improve over time, consistent with other findings about the cone of uncertainty. [9,10,11]

2.3 Distribution of Story Point Size

The distribution of story points had a P90/P10 ratio of 2.9 which again had characteristics of a lognormal or Weibull distribution. Although this was the distribution from the data, if a team wanted to narrow their distribution they could easily use story splitting techniques to obtain stories with more homogeneous sizes.

2.4 Decisions

While our findings support that project tracking is not significantly improved by using story point estimates, there may still be times where estimates are valuable or necessary. We looked at many of the business decisions that are encountered in software projects to examine the value of estimation. Practitioners should understand what decisions they are making and the implications that estimating or not estimating might have on those decisions. Our data along with the simulations provide a valuable context for those discussions.

3. Estimation in agile projects

Estimation in software development can mean many things. Teams estimate cost, effort, schedule or

value at varying degrees of detail and time horizons. Agile teams generally work with some form of user stories and/or tasks. When teams use estimates, they frequently work with estimates at the levels of granularity listed in Table 1.

Release	A release is a collection of features/stories that completes a delivery of interest to the customer. Estimates are typically in the form of effort, time and cost.
Feature or Epic	A high level story covering a broad area. It will generally be broken up into smaller stories. Estimates are typically representative of effort, and as story points, T-shirt sizes, or other.
Story	A unit of deliverable value to the customer. Stories will generally adhere to the INVEST [12] principles (Independent, Negotiable, Valuable, Estimable, Small, Testable). Typically estimates are in story points.
Task	Specific activities that the delivery team will do in order to complete a story. Typically estimated in hours.

Table 1

Our data are at the story level (so stories and story points), likewise our simulations were at that level for comparison. In Section 6 we suggest what our findings at the story level may imply regarding the value of estimates at other levels based on the decisions that the team is trying to address.

4. Analysis of the Data

Each project reported a tabulation of story points and number of stories delivered for each (timed) iteration. We initially focused on the decision area of project tracking, specifically comparing the predictive power of velocity with that of throughput.

4.1 Project Burnup Charts

In order to compare projects we rescaled the data for each project based on the overall time, story points and stories for the release. With this rescaling the burnup chart for each project starts at the origin and ends at the upper-right corner of the unit square. In theory, an agile project with “perfect” estimation would have a straight line from (0,0) to (1,1).

Figure 2 shows a plot of story points and stories versus time for the first three projects and the

aggregate curve of all the projects. Our first impression is that the trends for story points and stories are nearly identical. We also note from the Aggregate curve that many projects complete all stories by time $t=0.9$, but require additional time at zero velocity to finally be “complete.”

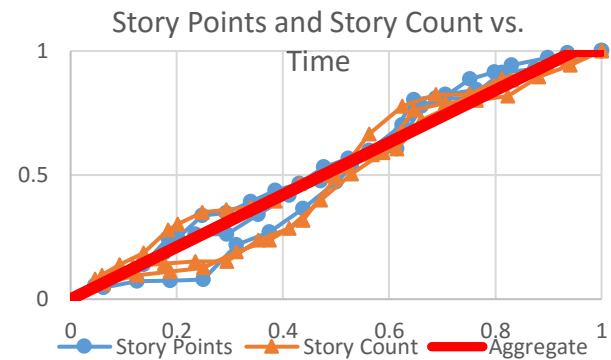


Figure 2

4.2 Velocity and Uncertainty Range over Time

For all projects, for each iteration we also calculated the rescaled velocity. Figure 3 shows the P90 and P10 bands of velocity over time and Table 2 shows the numerical values.

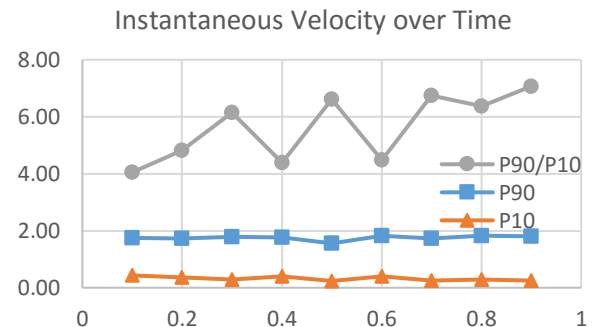


Figure 3

Time	P90	P10	P90/P10
0.1	1.75	0.43	4.06
0.2	1.72	0.36	4.81
0.3	1.79	0.29	6.14
0.4	1.76	0.40	4.39
0.5	1.56	0.24	6.62
0.6	1.82	0.41	4.49
0.7	1.73	0.26	6.73
0.8	1.83	0.29	6.38
0.9	1.81	0.26	7.07

Table 2

The range of the P90/P10 is between 4.06 and 7.07 with an overall P90/P10 of 5.1. We observe that the range is moderately consistent and does not reduce over time, in fact it may actually get worse.

4.3 Project Projections Using Velocity

To determine the predictive power of using velocity versus using throughput, at each iteration we made projections based on the remaining work using the average velocity and throughput. We compare these projections against the known actuals to determine the relative errors. Figure 4 shows the burnup chart for all projects plotted on the same axis scale. In this particular case we have data through $t=0.3$. Then using the velocity thus far we project out the anticipated release date. The error in the projection is the delta from 1.0. On this same figure we show the P90 and the P10 projections. To determine the actual error bands we subtract out the current projection time ($t=0.3$). For this chart (with approximations to simplify the math) we have P90 of 1.5, and P10 of 0.6. The revised P90 and P10 are 1.2 and 0.3 respectively, which gives a P90/P10 ratio of 4.0.

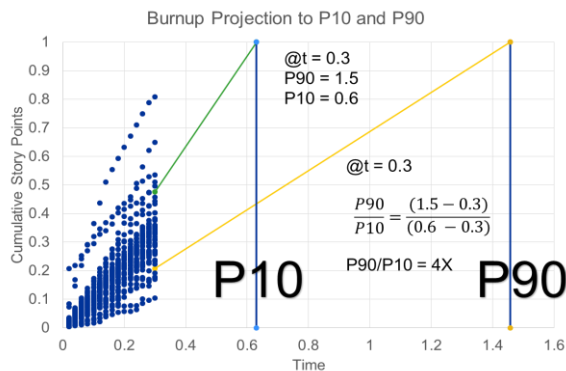


Figure 4

Figure 5 shows the results of the range of errors in the projections made as a function of time. The ratio of the throughput to velocity P90/P10 ratios averages 0.94 indicating throughput to be a better predictor with a 6% narrower error range. But there is really no significant difference in the accuracy of the forecast projections. While both velocity and throughput projections are similar in their degree of accuracy, neither is a phenomenal predictor with a P90/P10 ratio of about 3.5 in both situations.

In practical terms, whether using velocity or throughput, if a team forecasts that they have about 6 months remaining, the P10 to P90 bands for 80% confidence are roughly 3.2 to 11.2 months. This does

not bode well for teams or stakeholders that are expecting estimates that are commitments or even within 25% accurate as suggested by [13]. This implies a mismatch between the degree of accuracy expected and the reality of the range of uncertainty that is generally encountered. The uncertainty range from this data is consistent with other research [9,10,11].

Projection of P90/P10 ratios for Velocity and Throughput

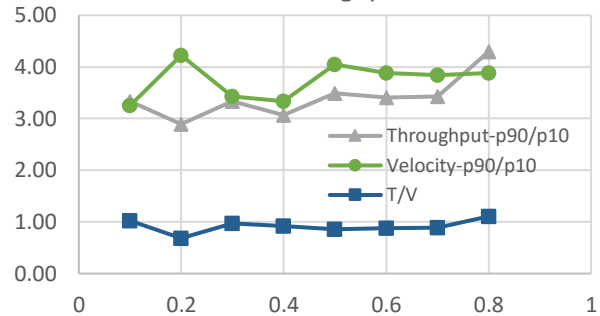


Figure 5

5. Analysis using Simulations

In order to better understand the findings from the data we simulated the results using a Monte Carlo technique.

5.1 Simulation Approach

For each scenario, we simulated 1000 projects each with 50 stories. For each story we sampled the story point distribution to establish its nominal story points. If a bucketing approach was used then we determined how it would be "bucketed." This was used for tabulating story points. Using the original (pre-bucketed) story points, the actual time was derived from the estimation accuracy distribution. As with the analyzed data, we converted to rescaled values for consistency across projects.

5.2 Simulation Parameters

The parameters that we explored are fourfold:

Story Point Distribution: We started with the empirical distribution from the data and then also explored distributions like Weibull and lognormal curve fits and variations on these distributions to see the impact that it would have on the simulations. Our baseline was Weibull and lognormal distributions with a P90/P10 ratio of 3.0. Story point distribution is one area where individual teams have significant control

over how they define and split their stories. Some teams may have large epics interspersed with smaller stories, while others may take the effort to split stories into relatively similar sizes.

Estimation Accuracy: In addition to the empirical distribution from the velocity data, we also used curve fits for lognormal and Weibull, with P90/P10 ratios of 5.2 and 5.7 respectively. In further scenarios we varied the P90/P10 ratio to explore its sensitivity. The data also showed a correlation between Story Point Distribution and Estimation Accuracy which we used in our simulations.

Bucketing Approach: Agile teams often use some form of bucketing for story points. Planning Poker [14] is a popular approach using a modified Fibonacci series of { 1/2, 1, 2, 3, 5, 8, 13, 20, 40, 100}. Other approaches include powers of 2 or 4. We varied the bucketing to find the impact they might have on the overall predictive power.

Hardening Effort: Roughly 50% of the projects required 2-12% of the total time at the end with zero velocity. We included this in the simulation as well.

5.3 Baseline Simulation Analysis

Table 3 shows the analyses of the different story point distributions, accuracy distributions, and impact of bucketing. For each scenario we show the P90/P10 ratio of the projections using velocity and throughput respectively. We also show the percentage improvement in range reduction by using velocity. A negative improvement would indicate that throughput was better than velocity.

Story Point #	Dist.	P90/p10	Estimation Accuracy	P90/p10	Bucketing	p90/p10 T-put	p90/p10 Velocity	Velocity %Impr
1	Empirical	2.9	Empirical	4.5	None	4.04	3.63	11%
2	Lognorm	3.0	Lognorm	5.7	None	2.60	2.55	2%
3	Weibull	3.0	Weibull	5.2	None	3.34	3.33	0%
4	Lognorm	3.0	Weibull	5.2	None	3.37	3.29	2%
5	Lognorm	3.0	Weibull	2.5	None	1.86	1.86	0%
6	Lognorm	3.0	Weibull	6.0	None	3.87	3.76	3%
7	Lognorm	2.0	Weibull	5.2	None	3.18	3.22	-1%
8	Lognorm	6.0	Weibull	5.2	None	3.95	3.52	12%
9	Lognorm	3.0	Weibull	5.2	None	3.37	3.29	2%
10	Lognorm	3.0	Weibull	5.2	Fibonacci	3.37	3.38	0%
11	Lognorm	3.0	Weibull	5.2	2x	3.37	3.50	-4%
12	Lognorm	3.0	Weibull	5.2	3x	3.37	3.63	-7%
13	Lognorm	3.0	Weibull	5.2	4x	3.37	3.61	-7%

Table 3

The first three scenarios show the results of using various story point and accuracy estimation distributions to see how well we could match the data from this study. We started with the empirical curves and also tried lognormal and Weibull curve fits. We believe the curve fits smooth some of the noise in the empirical data resulting in a better match. Scenario 4 using a lognormal fit for story point distribution and a Weibull fit for estimation accuracy gave a good match with our observed data. The plot of the P90/P10 projection ratios are all near 3.5 as shown in Figure 6.

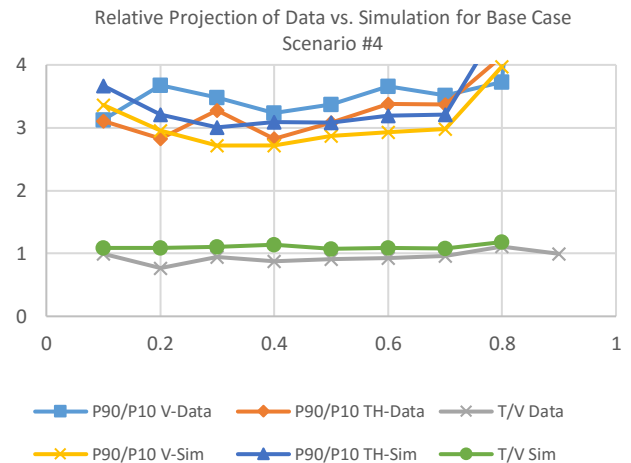


Figure 6

The simulations showed projections using velocity to be nearly identical to using throughput, similar to what we found in the data. This gave us confidence to proceed with some sensitivity analysis.

5.4 Sensitivity Analysis

Estimation Accuracy: For scenarios #5 and #6 we modified the estimation accuracy distribution to P90/P10=2.5 and 6.0 respectively. A surprising finding is that improved estimation accuracy helps both velocity and throughput projections equally. If estimation accuracy range is narrower, then the corresponding distribution of actual story points is also narrower.

Story Point Distribution: We then explored the range of the story point distributions varying the baseline lognormal distribution in scenarios #7 and #8. There is a clear relationship: the broad P90/P10 ratio of 6.0 gives much more benefit to velocity (11%), while the very narrow P90/P10 of 2.0 shows throughput to be slightly better than velocity (-1%). The narrower story distribution shows improved

predictability over the broader distribution, due to the story distribution constraining the uncertainty.

Bucketing: Lastly scenarios 9-13 investigated the impact of bucketing: comparing no bucketing with Fibonacci, power of 2, power of 3 and power of 4 bucketing. As expected, the buckets do not impact throughput at all, but they do degrade the predictive power of velocity somewhat as buckets get large. The impact is not significant as the velocity advantage starts at 2% for no buckets and inverts to favor throughput (-7%) for large buckets.

5.5 Summary of Simulation Results

We were able to match the data quite well with our simulation which gave us confidence to explore scenarios. The advantage of using velocity is almost negligible except when story point distribution is large. This is good news as the team has a lot to say about how they split their stories so as to reduce the range of the story distribution.

6. Implications on Decisions

One of the primary reasons for estimating is to improve decision-making. What do our findings suggest regarding some key decisions? We list some of the key decisions that project teams, sponsors and customers typically make during a software project.

- Decisions at project sanction
 - What is the cost, is it worth doing?
 - When is the target delivery?
 - What is the critical scope?
 - Do we have the right investment?
- Decisions to steer towards the release
 - Are we on target to meet our commitments?
 - What are the scope/schedule tradeoffs?
 - Is it worth continuing?
 - Can we do anything to accelerate delivery?
 - What is the cost of delay?
- Decisions to help with managing iterations
 - Can we make our iteration commitment?
 - What is our capacity?

Let's look at each of these categories in the context of our findings from the data and the simulations:

Decision	Role of Estimation
Decisions to steer towards the release	We observed from both the data and the simulations that story point estimates provide minimal improvement in forecasting compared to using throughput. The simulations showed that estimates may help when there is a large range of story distribution, although an alternative approach would be to split large stories so that the overall distribution is not large. When estimating a container of mixed nuts, we don't really care too much whether we have smaller peanuts or larger brazil nuts, but we do want to spot any coconuts!
Decisions to help with managing iterations	Many teams use detailed task estimation to help them manage their iterations. We did not have access to task estimations for this study, however the findings with story points should be very enlightening. Task estimation can often be a very time consuming activity. Teams should take a hard look at how much value they are getting from these estimations.
Decisions at project sanction	Some level of macro-estimation of costs and benefits is likely necessary for business decisions. If the benefits are so overwhelming that it should be done at any cost, then it could be wasteful to spend time on estimating something that does not impact the decision. In general teams should be careful of the trap to spend too much time on cost estimation. In fact, a study of a number of projects at a major organization found that value generated was negatively correlated to cost forecast accuracy [15]. Too much emphasis on cost or on reduction of uncertainty can destroy forecasting accuracy of value predictions.

Table 4

7. Practitioner's Guide

7.1 Velocity vs. Throughput

With the typically observed story point distribution range, our results show there is minimal added value to using velocity over using throughput for estimating purposes. When story size distribution is very large, then there is some improvement gained from using velocity.

7.2 Hardening

About half the projects required between 2-12% of the overall timeline with zero velocity at the end of the project, most likely for hardening activities. Unless teams have reasons to believe that they will not require such activities, we recommend adding stories (and story points if used) for such activities.

7.3 Estimation Accuracy

This study provides additional confirmation that the range of uncertainty with software estimation accuracy is significant and we can confidently say that this range of uncertainty is much larger than many decision makers realize. An interesting finding was that improvements in estimation accuracy helped throughput projections just as much as velocity projections. So while improving estimation accuracy may be a noble goal it is not a reason to favor velocity over throughput.

7.4 Bucketing of Estimates

While there was some degradation of the predictive power of velocity as buckets get very large, the overall impact is still very small. Since one reason for bucketing approaches is to expedite estimation, this finding suggests that teams that chose to estimate may continue to use them. However, we have seen situations where religious adherence to bucketing approaches slowed down the estimation process and in those circumstances teams may be better suited with simpler approaches. Bucketing or #NoBucketing? You decide.

7.5 Uncertainty over Time

Perhaps a bit more bad news for teams and decision makers is that it doesn't get better over time. The range of relative uncertainty of the work left to be done is large and stays large over time, which is consistent with other findings [9,10,11].

7.6 Decisions

Decisions are being made at multiple levels. For some decisions there may be value to estimates of stories or story points. But those estimates most likely have very large uncertainty ranges. The important question for the team is to understand the decisions they care about, and to comprehend the range of uncertainty to make the appropriate decisions. Decision makers would be wise to learn more about making decisions under uncertainty. There is significant research in many other industries (e.g oil and gas exploration, financial institutions, actuaries, etc.)

7.7 Estimates or #NoEstimates

To paraphrase Polonius' advice to Laertes, "Neither an Estimator nor a #NoEstimation bigot be, for estimation oft implies a false sense of both accuracy and certainty, while NO estimates may make suboptimal decisions. To thine own self (and team) be true."

8. References

- [1] <http://zuill.us/WoodyZuill/2013/05/17/the-noestimates-hashtag/>
- [2] <http://guide.agilealliance.org/guide/velocity.html>
- [3] IFPUG (2012). The IFPUG Guide to IT and Software Measurement. Auerbach Publication
- [4] G. Dinwiddie, "Feel the Burn: Getting the Most Out of Burn Charts," Better Software, pp. 26-31, July/August 2009
- [5] T. Sulaiman, B. Barton, T. Blackburn, "AgileEVM - earned value management in Scrum Projects," AGILE 2006 (AGILE'06), Minneapolis, MN, 2006, pp. 10 pp.-16.
- [6] V. Duarte, *NoEstimates: How To Measure Project Progress Without Estimating*, OikosofySeries, 2016
- [7] <https://bitly.com/NoEstimatesProjectsDB>
- [8] OECD (2016), "Income inequality", in *OECD Factbook 2015-2016: Economic, Environmental and Social Statistics*, OECD Publishing, Paris.
- [9] T. E. Little, "Schedule Estimation and Uncertainty Surrounding the Cone of Uncertainty," IEEE Software, vol.23, no. 3, pp. 48-54, May/June 2006
- [10] J. L. Eveleens, C. Verhoef, "Quantifying IT forecast quality," *Science of Computer Programming*, vol. 74, no. 11-12, pp.934-988, Sept, 2009.
- [11] L. Cao, "Estimating Agile Software Project Effort: An Empirical Study", (2008). *AMCIS 2008 Proceedings*. Paper 401.
- [12] W. Wake, "INVEST in Good Stories, and SMART Tasks", <http://xp123.com/articles/invest-in-good-stories-and-smart-tasks/>
- [13] S.D. Conte, H.E. Dunsmore, V.Y. Shen, *Software Engineering Metrics and Models*, Menlo Park, CA: Benjamin/Cummings Pub. Co., 1986.
- [14] https://en.wikipedia.org/wiki/Planning_poker
- [15] J.L. Eveleens, M. van der Pas, C. Verhoef, "Quantifying forecast quality of IT business value", *Science of Computer Programming*, vol. 77, no. 3, pp. 314-354, Mar. 2012