# The Rise and Fall of the Chaos Report Figures

**J. Laurenz Eveleens and Chris Verhoef,** *Vrije Universiteit Amsterdam*

Although the Standish Group's Chaos reports are often used to indicate problems in application software development project management, the reports contain major flaws.

For many years, researchers and practitioners have analyzed how to successfully manage IT projects. Among them is the Standish Group, which regularly publishes its findings in its Chaos reports. In 1994, Standish reported a shocking 16 percent project success rate, another 53 percent of the projects were challenged, and 31 percent failed outright.[1] In subsequent reports Standish updated its findings, yet the figures remained troublesome. These reports, derived from the Standish Group's longitudinal data, suggest that many efforts and best practices to improve project management hardly help increase project success. Over the years, their figures have attracted tremendous attention.

However, we question the validity of their figures. Robert Glass[2,3] and Magne Jørgensen and his colleagues[4] indicated that the only way to assess the Chaos results' credibility is to use Standish's data and reiterate their analyses. But there's another way: obtain your own data and reproduce Standish's research to assess its validity. We applied the Standish definitions to our extensive data consisting of 5,457 forecasts of 1,211 real-world projects totaling hundreds of millions of euros. Our research shows that the Standish definitions of *successful* and *challenged* projects have four major problems: they're misleading, one-sided, pervert the estimation practice, and result in meaningless figures.

## Misleading Definitions

The Standish Group published the first Chaos report in 1994, which summarized Standish's research findings and aimed to investigate causes of software project failure and find key ways to reduce such failures.[1] The group also intended to identify the scope of software project failures by defining three project categories that we recall verbatim:

- *Resolution Type 1, or project success.* The project is completed on time and on budget, offering all features and functions as initially specified.
- *Resolution Type 2, or project challenged.* The project is completed and operational but over budget and over the time estimate, and offers fewer features and functions than originally specified.
- *Resolution Type 3*, or project impaired. The project is cancelled at some point during the development cycle.[1]

To find answers to their research questions, Standish sent out questionnaires. Their total sample size was 365 respondents representing 8,380 applications. On the basis of the responses, Standish published overall percentages for each project cat-


0740-7459/10/$26.00 © 2010 IEEE

egory. Standish updated its figures in subsequent years (see Table 1). A number of authors published these figures in various white papers.[1,5–7]

The figures indicate large problems with software engineering projects and have had an enormous impact on application software development. They suggest that the many efforts and best practices put forward to improve how companies develop software are hardly successful. Scientific articles and media reports widely cite these numbers. Many authors use the figures to show that software development project management is in a crisis. The numbers even found their way to a report for the President of the United States to substantiate the claim that US software products and processes are inadequate.[8]

The figures' impact and their widespread use indicate that thousands of authors have accepted the Standish findings. They're perceived as impeccable and unquestionable. However, the Standish definitions of successful and challenged projects are problematic. Standish defines a successful project solely by adherence to an initial forecast of cost, time, and functionality. The latter is defined only by the amount of features and functions, not functionality itself. Indeed, Standish discussed this in its report: "For challenged projects, more than a quarter were completed with only 25 percent to 49 percent of originally specified features and functions."[1]

So, Standish defines a project as a success based on how well it did with respect to its original estimates of the amount of cost, time, and functionality. Therefore, the Standish "successful" and "challenged" definitions are equivalent to the following:

- *Resolution Type 1, or project success.* The project is completed, the forecast to actual ratios (*fla*) of cost and time are ≥1, and the *fla* ratio of the amount of functionality is ≤1.
- *Resolution Type 2, or project challenged.* The project is completed and operational, but *fla* < 1 for cost and time and *fla* > 1 for the amount of functionality.

The reformulated definitions illustrate that the definitions are only about estimation deviation.

Jørgensen and his colleagues show that the definitions don't cover all possibilities.[4] For instance, a project that's within budget and time but that has less functionality doesn't fit any category. In this article, we assume a project that doesn't comply with one or more of the success criteria belongs to the challenged-project category.

Standish calculates its success measure by counting the number of projects that have an initial forecast larger than the actual for cost and time, and one that's smaller for functionality. This is divided by the total number of projects to calculate the success rates. Standish Group defines its success measure as a measure of estimation accuracy of cost, time, and functionality.

**Table 1**

**Standish project benchmarks over the years**

| Year | Successful (%) | Challenged (%) | Failed (%) |
|---|---|---|---|
| 1994 | 16 | 53 | 31 |
| 1996 | 27 | 33 | 40 |
| 1998 | 26 | 46 | 28 |
| 2000 | 28 | 49 | 23 |
| 2004 | 29 | 53 | 18 |
| 2006 | 35 | 46 | 19 |
| 2009 | 32 | 44 | 24 |

In reality, the part of a project's success that's related to estimation deviation is highly context-dependent. In some contexts, 25 percent estimation error does no harm and doesn't impact what we would normally consider project success. In other contexts, only 5 percent overrun would cause much harm and make the project challenged. In that sense, there's no way around including more context (or totally different definitions) when assessing successful and challenged projects. However, the Standish definitions don't consider a software development project's context, such as usefulness, profit, and user satisfaction.

This illustrates the first problem with the definitions. They're misleading because they're solely based on estimation accuracy of cost, time, and functionality. But Standish labels projects as successful or challenged, suggesting much more than deviations from their original estimates.

## Unrealistic Rates

The next issue is whether the Standish estimation accuracy definitions are sound. They are not. The Standish Group's measures are one-sided because they neglect underruns for cost and time and overruns for the amount of functionality.

We assessed estimation accuracy with two tools. We derived the first from Barry Boehm's now-famous cone of uncertainty, a plot that depicts forecast to actual ratios against project progression.[9] This plot shows how the forecasts are made, what deviations they contain, and whether institutional biases exist.

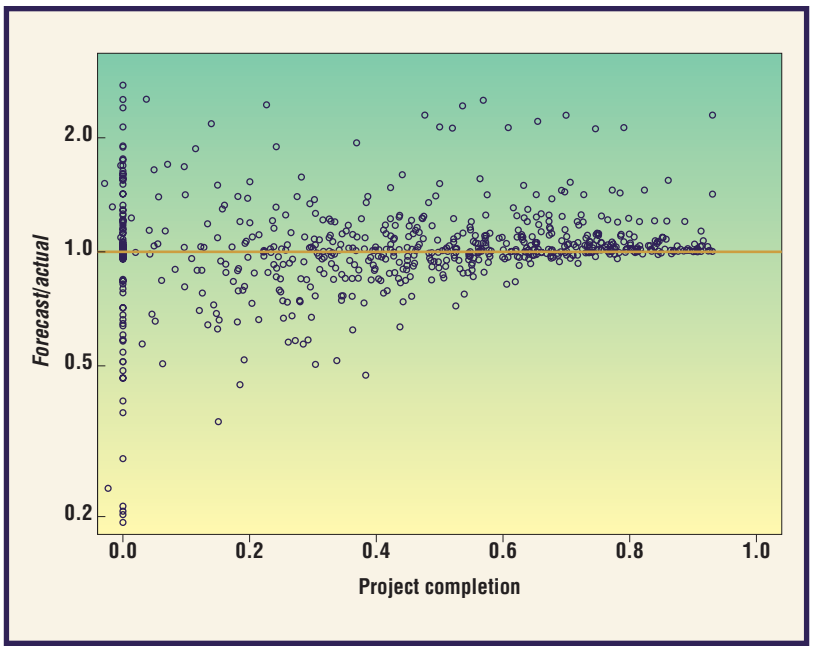


**Figure 1. 667 *f/a* ratios for 140 project costs of organization Y, where *f* is forecast and *a* is actual. The ratios are spread equally below and above the horizontal line *f/a* = 1, indicating the forecasts are unbiased. The ratios also show that the quality of the forecasts is high compared to the literature.[10,11]**

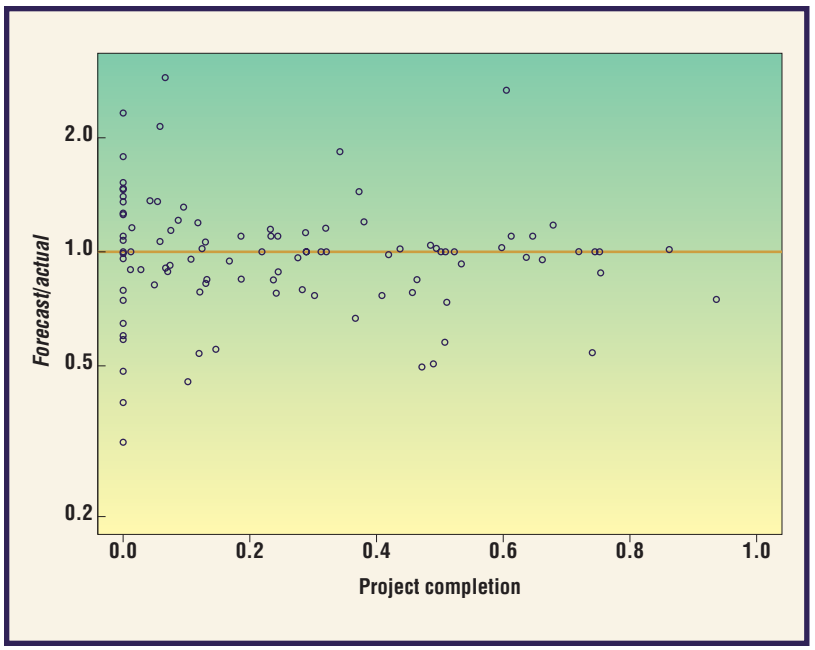


**Figure 2. 100 *f/a* ratios for 83 project function points of organization Y, where *f* is forecast and *a* is actual. The ratios are close to and centered around the horizontal line. This indicates the forecasts are unbiased and of high quality.**

The second is Tom DeMarco's Estimation Quality Factor (EQF), a time-weighted estimation accuracy measure he proposed in 1982.[10] The higher a forecast's EQF value, the higher its quality. An EQF value of 5 means the time-weighted forecasts of a single project deviate on average 1/5, or 20 percent, from the actual.

We applied Boehm's and DeMarco's work to our own data and detected large biases that the organizations weren't aware of. We introduce two data sets from an anonymous multinational corporation to prove that the one-sided Standish definitions lead to unrealistic rates.

## Cost

The first case study concerns a large financial-services provider. From this organization, Y, we obtained data on 140 software development projects conducted from 2004 to 2006. The organization made 667 forecasts for these projects' total costs. We divided the forecasted cost with the actual project cost and plotted the ratios as shown in Figure 1. The horizontal axis represents project progression. The figure depicts the start of a project at zero and represents project completion by 1.0. The vertical axis shows the *f/a* ratio's value. For instance, a data point at project completion 0.2 and an *f/a* ratio of 2 indicates a forecast was made when the project was one-fifth completed. This forecast was two times the actual, meaning the project turned out to be 50 percent of the estimated cost.

The *f/a* ratios in Figure 1 resemble Boehm's conical shape, with the forecasts centered around the actual value. A median *f/a* ratio of 1.0 supports this finding. The forecasts' quality is relatively high, with a median EQF value of 8.5. This indicates that half the projects have a time-weighted average deviation of 12 percent or less from the actual. Compared to results from the literature, this organization makes best-in-class forecasts.[10,11]

It turned out that an independent metrics group assessed this organization's forecasts. This group made its own cost calculations next to those of the project managers. If large discrepancies arose, these needed to be resolved before any budget was approved. This caused forecasts to aim at predicting the actual value. Yet, even though this organization's cost forecasts are accurate, when we apply the Standish definitions to the initial forecasts, we find only a 59 percent success rate.

## Functionality

From the same organization Y, we obtained data for 83 software development projects from 2003 to 2005. In total, the organization's estimators made 100 forecasts for the projects' functionality, calculated in function points.[12]

The functionality *f/a* plot in Figure 2 shows a situation similar to the *f/a* ratios for the costs. The bias is negligible based on the figure and a median *f/a* ratio of 1.0. Except for some outliers, the *f/a* ratios converge to the actual value. The functionality forecasts have a median EQF of 6.4. This means that the function-point forecasts of half the projects have a time-weighted average deviation of 16 percent or less from the actual amount.

Multiple experienced function-point counters calculated the projects' functionality. Because they weren't involved with the projects' execution, their only incentive was to predict the actual value. However, despite the forecasts' accuracy, when we apply the Standish definitions to the initial forecasts, we find only a 55 percent success rate.

### Combined

Fifty-five software development projects contained forecasts and actuals of both cost and functionality. There were 231 cost forecasts and 69 functionality forecasts. Both cost and functionality forecasts were unbiased and converged to the actual value. The median EQF for the cost forecasts is 9.0; for the functionality forecasts, it's 5.0. So, half the projects have a time-weighted average deviation of 11 percent for cost and 20 percent deviation for functionality.

We applied the reformulated Standish definitions to the initial forecasts of the combined data. Even without taking into account failed projects and the time dimension, the best-in-class organization Y obtains a success rate of 35 percent. Yet, the median EQF of both initial forecasts of costs and functionality is 6.5, showing that half the projects have an average time-weighted deviation of only 15 percent from the actuals. If this organization is already so unsuccessful in two dimensions according to Standish, it's hardly surprising that Standish found only a 16 percent success rate in its first report.[1]

These case studies show that organization Y obtains unrealistically low success rates for the individual cost and functionality forecasts owing to the definitions' one-sidedness. Combining these already low rates further degrades the success rate. Clearly, the Standish success rates don't give an accurate indication of true estimation accuracy of cost and functionality in the case of an unbiased best-in-class organization.

## Perverting Accuracy

The third problem is that steering on the Standish definitions causes large cost and time overestimations (and large functionality underestimations), which perverts rather than improves estimation accuracy.

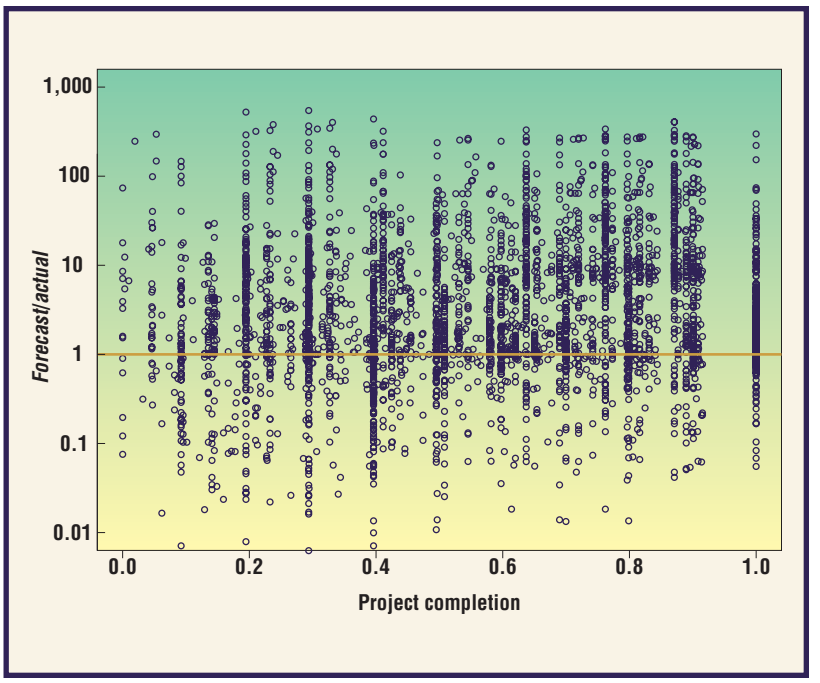


**Figure 3. 3767 *f/a* ratios for 867 project costs of organization X, where *f* is forecast and *a* is actual. The forecasts show large deviations and do not converge to the actuals over time. The figure shows that these forecasts are generally overestimated and of low quality.**

We obtained data from a large multinational organization, X, comprising 867 IT-intensive projects that it began and completed in 2005 or 2006. In total, the organization made 3,767 forecasts of the projects' costs.

The *f/a* ratios in Figure 3 show that the organization's forecasts were generally higher than the actuals. Also, the data doesn't show a conical shape as we'd expect from Boehm's cone of uncertainty. Projects even had surplus budget after completion. After discussion with the organization, we found it steered on Standish project success indicators. The organization adopted the Standish definitions to establish when projects were successful. This caused project managers to overstate budget requests to increase the safety margin for success. However, this practice perverted the forecasts' quality, making it low with a median EQF of 0.43. So, 50 percent of the projects have a time-weighted average deviation of 233 percent or more from the actual.

## Meaningless Figures

The fourth major problem is that the Standish figures are meaningless. Organization X showed that large biases occur in practice. Even if a company doesn't steer on Standish's key

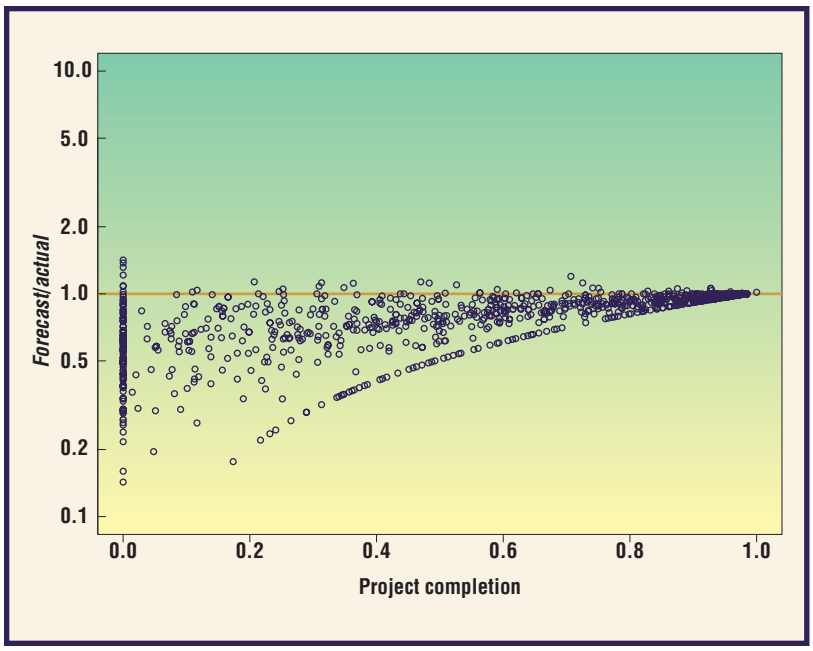


**Figure 4. 923 *f/a* ratios for 121 project durations of Landmark Graphics, where *f* is forecast and *a* is actual. The forecasts are reasonably close to the horizontal line, yet, most *f/a* ratios are below it. The figure indicates the forecasts are biased toward underestimation.**

performance indicators, biases exist. We show this by introducing another case study from an earlier *IEEE Software* paper.[13] Comparing all the case studies together, we show that without taking forecasting biases into account, it's almost impossible to make any general statement about estimation accuracy across institutional boundaries.

## Time

Landmark Graphics is a commercial software vendor for oil and gas exploration and production. We obtained data from Todd Little of Landmark Graphics, which he reported in *IEEE Software*,[13] consisting of 121 software development projects carried out from 1999 to 2002. Little provided 923 distinct forecasts that predict these 121 projects' duration. We performed the same analysis as before by plotting the forecast to actual ratios (see Figure 4).

Most forecasts this organization made are lower than the actual. So, projects take longer than initially anticipated. The median EQF is 4.7. This means that half the projects have a time-weighted average deviation from their forecasts of 21 percent or less from the actual. Landmark Graphics' institutional bias was to forecast the minimum value instead of the actual value. This caused most forecasts to be lower than the actuals.

## Applying Standish's Definitions

In two of the three organizations, the forecasts were significantly biased. With organization Y, we determined that the institutional bias was negligible. In organization X, the forecasts were much higher than the actual values because estimators took large safety margins into account. With Landmark Graphics, most forecasts were lower than the actual values because the company predicted the minimal time required to finish the project.

To illustrate how forecasting biases introduced by different underlying estimation processes affect the Chaos report figures, we applied Standish's definitions to all the cases. Because Standish deals with initial forecasts, we also used the initial forecast of each project. This is a subset of all data points shown in the *f/a* plots in Figures 1–4.

Also, our resulting figures are an upper bound for the Chaos successful-project figures. First, our figures don't incorporate failed projects. If we took failed projects into account, our case studies' success rates would always be equal to or lower than the current percentages.

Second, in each case study, we present only cost, time, or functionality data, except in one instance where we present both cost and functionality. In our analysis, we assume that the remaining dimensions are 100 percent successful, meaning our percentages are influenced by only one or two dimensions. If data for all three dimensions (cost, time, and functionality) is available and taken into account, the success rates will always be equal to or lower than the successful percentages calculated for only one or two dimensions. Still, these rates suffice to prove that Standish's success and challenge rates don't reflect the reality.

Table 2 shows the numbers calculated according to Standish's definitions for our case studies along with those of a fictitious organization having the opposite bias of Landmark Graphics. The table provides an interesting insight into the Standish figures. Organization X is very successful compared to the other case studies. Nearly 70 percent of the projects are successful according to the Standish definitions. On the other end, Landmark Graphics has only 6 percent success. Organization Y is in-between with 59 percent success for costs, 55 percent success for functionality, and 35 percent success for both.

However, the *f/a* plots and their median EQFs clearly show that this is far from reality. Landmark Graphics' and organization Y's initial fore-

**Table 2**

**Comparing Standish success to real estimation accuracy**

| Source | Successful (%) | Challenged (%) | Median estimation quality factor of initial forecasts |
|---|---|---|---|
| Organization X | 67 | 33 | 1.1 |
| Landmark Graphics | 5.8 | 94.2 | 2.3 |
| Organization Y cost | 59 | 41 | 6.4 |
| Organization Y functionality | 55 | 45 | 5.7 |
| Organization Y combined | 35 | 65 | 6.5 |
| 1/Landmark Graphics | 94.2 | 5.8 | 2.3 |

casts deviate much less from their actuals than in the case of organization X, which overestimates from tenfold to a hundredfold, as Figure 3 shows. Also, the other organizations' estimation quality outperforms organization X, which the median EQF of their initial forecasts illustrates: 2.3 for Landmark Graphics, 6.4 for organization Y's costs and 5.7 for organization Y's functionality, versus 1.1 for organization X. So, half of Landmark Graphics' initial forecasts deviate only 43 percent from the actual value, 16 percent for organization Y's costs and 18 percent for organization Y's functionality, versus 91 percent for organization X. Still, Standish considers organization X highly successful compared to the other organizations.

To further illustrate how easy it is to become highly successful in Standish's terms, we also presented 1/Landmark Graphics. This fictitious organization represents the opposite of Landmark Graphics. That is, the deviations to the actuals remain the same, but an overrun becomes an underrun and vice versa. Suddenly, 1/Landmark Graphics becomes highly successful with a 94 percent success rate. So, with the opposite institutional bias, Landmark Graphics would improve its Standish success rate from 6 percent to 94 percent.

These case studies show that the Standish figures for individual organizations don't reflect reality and are highly influenced by forecasting biases. Because the underlying data has an unknown bias, any aggregation of that data is unreliable and meaningless.

The influence of biased forecasts on the Standish figures isn't just evident from our figures. Standish's Chairman Jim Johnson clearly indicates that manipulating the figures is easy:

> *In 1998, they [the respondents] had changed their [estimating] process so that they were then taking their best estimate, and then doubling it and adding half again.*[14]

Johnson made this statement with respect to the drop in the reported average cost overruns between 1996 (142 percent) and 1998 (69 percent). In the article, Johnson says that he doesn't believe this change of process is the cause of the drop. However, our case studies show that forecasting biases have a giant influence on such figures. So, we believe that the change in the estimating process is most likely the cause of the drop in the reported cost overruns.

We developed methods based on Boehm and DeMarco's work that mathematically account for forecasting biases.[15] Our other paper contains more information about the case studies in addition to another one (totaling 1,824 projects, 12,287 forecasts, and 1,059+ million euros).[15] We propose bandwidths surrounding the actual value to determine whether forecasts are accurate. These bandwidths show that projects with relatively small underruns or overruns have accurate forecasts, whereas projects with relative large underruns or overruns have inaccurate forecasts. The mathematical implications are manifold and are out of the scope of this paper. But, we were able to derive figures that were exactly in line with the reality of our case studies. We hope that Standish will adopt our proposed definitions and methods for the rise and resurrection of their reports.

By ignoring the potential bias and forecasting quality, the figures of the Standish Group don't adequately indicate what, according to their definitions, constitutes a successful or challenged project. Some organizations tend to overestimate while others underestimate, so their success and challenge rates are meaningless because Standish doesn't account for these clearly present biases.

## About the Authors

**J. Laurenz Eveleens** is a PhD student at Vrije Universiteit Amsterdam's Department of Computer Science. His current research is aimed at quantifying the quality of IT forecasts. Eveleens has an MSc in business mathematics and informatics from VU University Amsterdam. Contact him at laurenz@few.vu.nl.

**Chris Verhoef** is a computer science professor at Vrije Universiteit Amsterdam and is a scientific advisor with IT-Innovator Info Support. His research interests are IT governance, IT economics, and software engineering, maintenance, renovation, and architecture. He has been an industrial consultant in several software-intensive areas, notably hardware manufacturing, telecommunications, finance, government, defense, and large service providers. Verhoef has a PhD in mathematics and computer science from the University of Amsterdam. He's an executive board member of the IEEE Computer Society Technical Council on Software Engineering and the vice chair of conferences. Contact him at x@cs.vu.nl.

This article isn't the first to challenge the Chaos report figures' credibility; a number of authors also "questioned the unquestionable."[2–4,16]

For instance, Nicholas Zvegintzov placed low reliability on information where researchers keep the actual data and data sources hidden.[16] He argued that because Standish hasn't explained, for instance, how it chose the organizations it surveyed, what survey questions it asked, or how many good responses it received, there's little to believe.

Also, Glass[2,3] felt the figures don't represent reality. Without plenty of successful software projects, he asserted, the current computer age would be impossible.

Moreover, Jørgensen and his colleagues expressed doubt about the numbers.[4] They unveiled a number of issues with Standish's definitions and argue that the resulting figures are therefore unusable. For instance, they argued that the definitions of successful and challenged projects focus on overruns and discard underruns.

Despite the valid questions our predecessors raised, no one had previously been able to definitely refute the Standish figures' credibility. Our research shows that Standish's definitions suffer from four major problems that undermine their figures' validity.

We communicated our findings[15] to the Standish Group, and Chairman Johnson replied: "All data and information in the Chaos reports and all Standish reports should be considered Standish opinion and the reader bears all risk in the use of this opinion."

We fully support this disclaimer, which to our knowledge was never stated in the Chaos reports.

## Acknowledgments

**This research received partial support from the Netherlands Organization for Scientific Research's Jacquard projects Equity and Symbiosis. We thank the anonymous reviewers and Nicholas Zvegintzov for commenting on this article.**

## References

1. *Chaos*, tech. report, Standish Group Int'l, 1994.
2. R. Glass, "IT Failure Rates—70% or 10–15%," *IEEE Software*, May 2005, pp. 110–112.
3. R. Glass, "The Standish Report: Does It Really Describe a Software Crisis?" *Comm. ACM*, vol. 49, no. 8, 2006, pp. 15–16.
4. M. Jørgensen and K. Moløkken, "How Large Are Software Cost Overruns? A Review of the 1994 Chaos Report," *Information and Software Technology*, vol. 48, no. 8, 2006, pp. 297–301.
5. D. Hartmann, "Interview: Jim Johnson of the Standish Group," 2006; www.infoq.com/articles/Interview-Johnson-Standish-CHAOS.
6. *Chaos: A Recipe for Success*, tech. report, Standish Group Int'l, 1999.
7. *Extreme Chaos*, tech. report, Standish Group Int'l, 2001.
8. B. Joy and K. Kennedy, *Information Technology Research: Investing in Our Future*, tech. report, President's Information Technology Advisory Committee, Feb. 1999.
9. B. Boehm, *Software Engineering Economics*, Prentice Hall, 1981.
10. T. DeMarco, *Controlling Software Projects*, Prentice Hall, 1982.
11. T. Lister, "Becoming a Better Estimator—An Introduction to Using the EQF Metric," www.stickyminds.com, 2002; www.stickyminds.com/s.asp?F=S3392_ART_2.
12. D. Garmus and D. Herron, *Function Point Analysis—Measurement Practices for Successful Software Projects*, Addison-Wesley, 2001.
13. T. Little, "Schedule Estimation and Uncertainty Surrounding the Cone of Uncertainty," *IEEE Software*, vol. 23, no. 3, 2006, pp. 48–54.
14. J. Johnson, "Standish: Why Were Project Failures Up and Cost Overruns Down in 1998?" InfoQ.com, 2006; www.infoq.com/articles/chaos-1998-failure-stats.
15. J.L. Eveleens and C. Verhoef, "Quantifying IT Forecast Quality," *Science of Computer Programming*, vol. 74, no. 11+12, 2009, pp. 934-988; www.cs.vu.nl/~x/cone/cone.pdf.
16. N. Zvegintzov, "Frequently Begged Questions and How to Answer Them," *IEEE Software*, vol. 20, no. 2, 1998, pp. 93–96.