

eXtreme Measurement:

Recognizing, understanding and avoiding measurement dysfunction

(or

'Why software measurement fails, and what to do about it')

C. C. Shelley

October 2009

Oxford Software Engineering
www.osel.co.uk
9 Spinners Court, 53 West End, Witney, Oxfordshire, England, OX28 1NH
shelley@osel.netkonect.co.uk

ABSTRACT

Software measurement should play a central role in software development and management, but it doesn't, yet. Its potential has not been realised; its development and the emergence of useful standards hindered, because of an important but poorly understood problem. Because people react in unexpected and sophisticated ways when they, their activities, their products, or their environment are measured the system of measurement can become dysfunctional, distorting the data or interfering with the measured. This dysfunction is usually recognized, or suspected, but lacking effective methods to deal with it is often dismissed with faux pragmatism, leading to the measures, and measurement, being marginalized. Measurement dysfunction follows a pattern that can be recognized and dealt with. This paper outlines a model of measurement dysfunction, analysing and explaining what happens, and why, and illustrated with examples from software development and other areas. A set of guidelines for recognizing dysfunction is presented, together with a set of radical action plans for avoiding or eliminating it.

1 Introduction

Good measurement delivers useful, often valuable information that can be critical to decision making and directing actions, but unsound measurement can mislead and confuse or distort decision making and even the artefacts or activities being measured.

This unsound measurement is often sensed by those affected by it, but are unsure what to do. Users of data may lack the confidence to make decisions based on it, rendering it valueless, and those charged with providing data, but having no use for it themselves, may provide data of dubious accuracy or fidelity.

This unsoundness is common, but elusive and rarely addressed, especially by measurement experts, who quite reasonably wish to promote the virtues of measurement and would not wish to high light its problems (or may not even be aware of them). But from time to time does receive some recognition:

'In my experience absolutely everyone who does it screws it up. So I think the concept's wrong. And I think it's kind of pointless to think, "Well, if they did it right, it would be OK." They don't do it right. The people who want to do it are inclined to do it wrong.'

Tom DeMarco

The problem is one of people's reactions to measurement, whether measuring or measured. Since software development is such a people intensive design process it is especially prone to measurement problems.

Robert D. Austin has described a compelling model of measurement dysfunction (Ref 1.). This will be

described below. But he has not described how it can be used. In part this may be because it implies limitations to the use of measurement that run counter to our cultural expectations of measurement and many may find difficult to accept.

If the model is reasonable then something has to be done to address the problems it identifies. Experiences with steps taken to minimize measurement dysfunction are described below.

2 A model of dysfunctional measurement

The way in which measurement has unintended effects has been known for a long time. It was described by Steven Kerr in 1975 (Ref 2.) who recognized that measured and actual performance deviate over time if the measures are inappropriate.

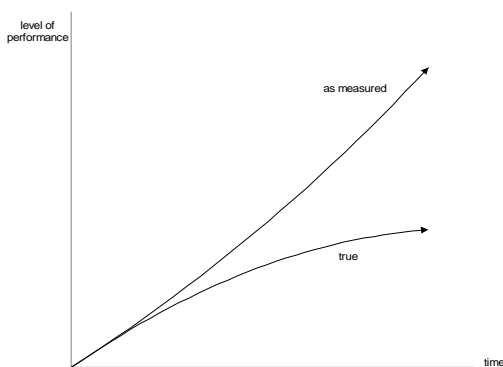


Figure 1.

Responses to this dysfunction are typically some form of faux pragmatism or denial:

- 'the measures we use are cost effective',
- 'they are a reasonable compromise',
- 'its the best we can do',
- 'what alternatives are there?'

which while sounding reasonable will trigger dysfunctional measurement and behaviour.

Before describing the model it is necessary to distinguish two fundamental types of measurement: *motivational* and *informational*. Motivational measurements are used to demonstrate: to measure outcomes (often as a basis for rewards), to encourage compliance with plans and are specifically intended to change that being measured. They are the basis for performance measures and performance improvement. Typically they include targets.

Informational measures are used to improve understanding: for day to day tactical use, for long term process improvement (to identify patterns and trends), and *should not* change that being measured. It is important to know how things really are.

Ideally these two types should be distinct. In practice they are difficult to segregate. The category a particular data set is assigned to depends not in on its character or use, but on its users. Where users of the data use it to motivate, or even when users *could use it* to motivate, even if they don't, then its value as informational measurement is lost.

Informational measurement is straightforward to use when measuring things, but if these things include, or are related to (self aware) people or groups (as is often the case in software development environments) who may believe that others will or may make judgements about them the measures become *de facto* motivational measures. Austin's model illustrates this:

Consider a (hypothetical) software team working for a manager or customer, testing a system and fixing the defects they find. (While considering this hypothetical team bear in mind how real people, including you, usually react.) The team runs tests, logs them, and any defects they find, and fixes the defects. They allocate effort to performing testing or fixing defects as they feel best matches the customers requirements.

The diagram below (Fig 2.) shows the context for this team's work. The dashed line indicates where a given amount of effort could be allocated: mostly to testing (bottom right), mostly to fixing (top left), or a mix. The best mix of effort tends to be a more or less equal mix of testing and fixing. Higher value is delivered by providing more effort, and the optimum mix tends to be delivered by this judicious mix of testing and fixing. This is represented by the solid value 'contour lines' where higher value is towards the top right of the diagram.

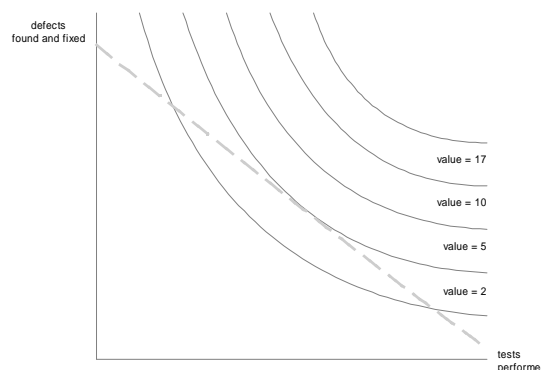


Figure 2.

The way the team allocates effort to testing and fixing is shown in Fig 3. Starting from a point of doing no testing or fixing (and with an expectant but

unsatisfied customer) the team move along the 'best mix' path toward the high value region until they reach their favoured point where they have satisfied the customer and need not expend further, unnecessary effort.

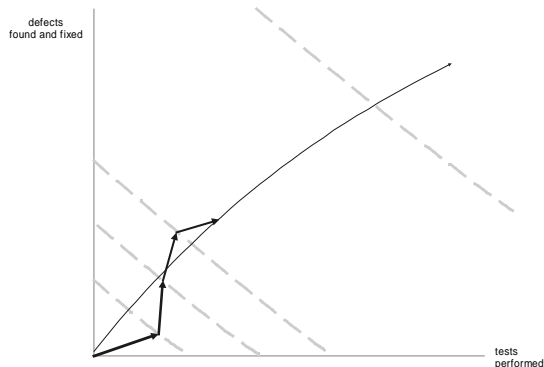


Figure 3.

This situation, where the team is *unsupervised* and allowed to allocate effort as they see fit, is satisfactory from the team's perspective, but not ideal from the customer's.

Now consider the situation where the customer can measure the teams allocations of effort to testing and fixing. The customer, quite reasonably, will want to get more for their money, and make payment contingent on higher levels of a effort being allocated to testing and fixing as shown in Fig 4:

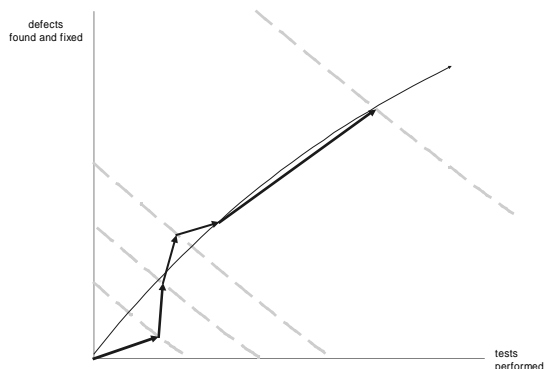


Figure 4.

Now with this *full supervision* the effort allocation has moved along the best mix path to deliver higher value. The customer is more satisfied. And perhaps the team may be less happy having to work harder to satisfy the customer and get paid. The overhead for this fully supervised situation is the cost of measurement, by the customer, of the critical dimensions of effort allocation of the team. If this measurement overhead is acceptable then more value is delivered (with more effort expended by the team). This presents a typical and attractive picture of the application of measurement.

In both instances (unsupervised or fully supervised) the use of measurement is functional and good, leading to the allocation of effort on the best mix path, albeit in different locations on that path.

Now consider what happens in this hypothetical team if one of the critical dimensions is not being measured – say defects found and fixed, so performance is assessed by the customer by tests effort, or tests completed. This is a not unrealistic situation. Many organizations use test schedules as the basis for managing this type of work. What happens in this situation of *partial supervision*?

It is human nature to expend effort on what will be (measured and) rewarded, and not on what is not rewarded. When test effort alone is measured this is where effort will be concentrated. Fig 5. shows that for a while value continues to increase as more tests are performed.

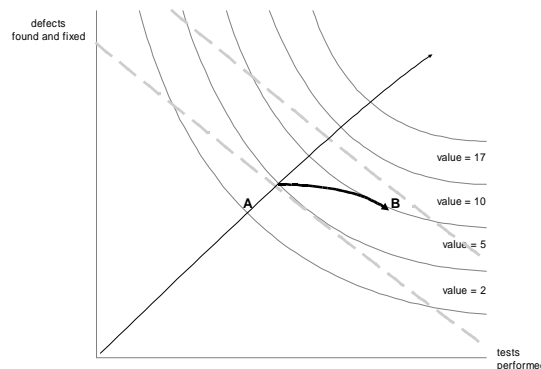


Figure 5.

But soon although the number of tests performed and test effort continues to increase, and the performance as measured by the customer continues to increase the value delivered to the customer declines (Fig 6.) as the team learns how to sustain or increase the number of tests performed with minimal (optimal from the teams perspective) effort. The measurement system has become dysfunctional.

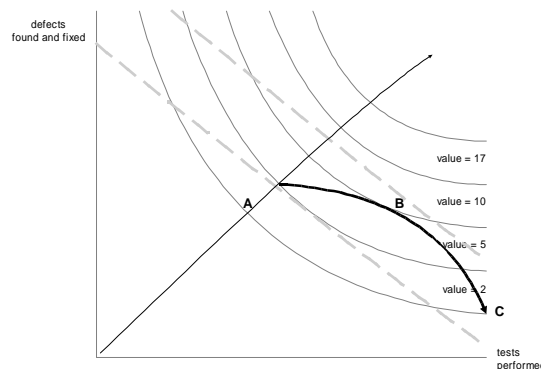


Figure 6.

This can be exacerbated if performance targets are introduced. The target acts as a buffer to limit performance and effort. The team will (again, quite reasonably) seek to reduce the effort they expend while maintaining the level of testing. From the teams perspective, improving their efficiency.

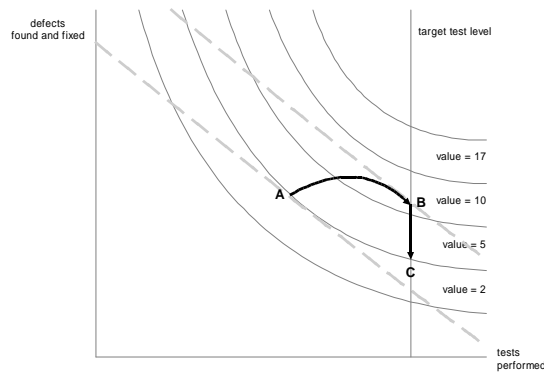


Figure 7.

In this way partial supervision – in this simplified case, meaning the customer uses testing only as the measure of performance – has distorted behaviour in a manner that reduces the value delivered.¹

In general partial supervision will, in time, distort behaviour in a way that reduces the value delivered, while, at the same time apparently improving performance, as discussed by Kerr.

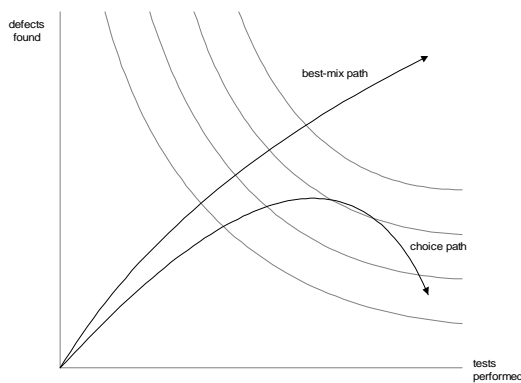


Figure 8.

The reasonable response when this is recognized is to require full supervision to ensure the best mix. This would indeed align value delivered with the measured performance but there are the problems of the cost of measurement which may become prohibitive and intrusive, and its feasibility. Consider the hypothetical team described above. To illustrate

¹In this case we have selected testing as the measure used to monitor and manage partial supervision. Consider what would happen if defects only were used as the measure. It takes software developers about 10 seconds to find ways of maximizing rewards for minimal effort!

the point it was assumed that two measures were adequate. In a real situation a number of other factors also have an important influence on the allocation of effort and the perception of value:

- effort to test
- time to test
- test effectiveness
- effort to fix
- time to fix
- definition of a defect ('its not a defect, its a feature')
- defect severity
- defect priority
- quality of fix
-other undiscovered factors?

Is it feasible *for the customer* measure all of these? Get any one wrong and dysfunction will result. If there are other undiscovered factors that are not measured dysfunction will result

Table 1 summarizes the limits of measurement.

	Unsupervised	Partially supervised	Fully supervised
Motivational		?	?
Informational	?		

Table 1.

Where measurement is used for informational purposes it is necessarily unsupervised . It becomes motivational if there is any element of supervision.

Where measurement is used to motivate it is necessarily supervised. This is acceptable where full supervision is possible; in simple or very well understood situations where things, not people, are measured. But in complex environments, involving people, like software organizations, partial supervision, whether performed knowingly or not,

whether performed with the best of intentions or not, will trigger dysfunction.

This has been sensed for some time. Deming is emphatic about the use of measurement and performance targets for both technical staff:

11a *“Eliminate numerical quotas for the work force... .. A quota is a fortress against improvement of quality and productivity. I have yet to see a quota that includes any trace of a system by which to help anyone do a better job. A quota is totally incompatible with never ending improvement. There are better ways.”*

W. Edwards Deming (Ref 3.)

and managers:

11b *“Eliminate numerical goals for people in management. Internal goals set in management of a company, without a method, are a burlesque... .. Focus on outcome is not an effective way to improve process or an activity... .. management by numerical goal is an attempt to manage without knowledge of what to do, and is in fact usually management by fear.”*

W. Edwards Deming

John Seddon also makes unambiguous statement about the use of measurement:

“Ownership of measures should be with the people who do the work.”

“Managers tend to think they ‘own’ the numbers. They should not.”

“People do what you count, not necessarily what counts.”

“Attention to output can increase costs.”

“Using measures for improvement starts with thinking differently.”

John Seddon (Ref 4.)

This can be difficult to accept. But Austin's model might make these assertions easier to understand and deal with. It appears that:

1. Informational measurement is fragile – and prone to become motivational, or to cease (in the author's experience).
2. Motivational measurement is prone to dysfunction – because full supervision is difficult, expensive (and undesirable?).
3. Because of this there is a limited 'information horizon' in complex, knowledge working environments beyond which

measurement data cannot reliably convey information.

4. But recognition of this information horizon can encourage and defend the practice of good informational measurement to aid software development and test optimization.

Sound measurement for informational purposes can be encouraged in a two step process:

1. Recognize dysfunctional measurement
2. Eliminate it.

The first step is reasonably straightforward, can be undertaken in most software environments, and can be extraordinarily informative. The second is more difficult requiring managers and measurement practitioners and others to accept limitations on the use of measurement, and requiring organizations to change their culture of information management and their expectations of 'visibility', 'transparency' and the ability to 'drill down'.

These two steps are described next.

3 Recognizing Dysfunction

To identify dysfunctional measurement an investigation of measurement activities is required. But before any investigation is begun permission is essential. Gain the consent of the persons who nominally own the data or are responsible for its collection, analysis and reporting. And ensure that information gained during the investigation is not, under any circumstances, attributable.

The objective is to identify measurement dysfunction (and, as useful side effects, measurement value (high or low), use (high or low) and its (sometimes considerable) cost too). Attitudes to measurement are also revealed. The approach of the investigation is to get answers to a series of questions, but not necessarily by asking the questions directly. Conduct confidential conversations and discussions with representatives at all levels within the organization. The questions you will ask are²:

1. What measurement data do you collect?
2. Why?
3. How do you use it?
4. What decisions or actions do you take based on your analysis of the data?

²Although it is not the answers that are important. It is what you learn on the way to getting the answers

In working to get answers to these questions you will gain an understanding of data ownership and use. Good, functional measurement will be indicated by a clear understanding of why the data is collected and what it is used for, often backed up by freely offered evidence - although it is sometimes difficult for those collecting and using measurement data well to articulate precisely why because its use seems self evident. In general good measurement has short 'lines of communication' or feedback loops with, perhaps the collector(s) also being the user(s). It will tend to be, necessarily, obscure and low profile. In contrast, low value or dysfunctional measurement will tend to be high profile, and common to all. It can be indicated by data collected but not used by the collector, with data being passed along lines of communication without use or clear understanding, being delivered in reports or summaries to nominal users that lack confidence in the data to make decisions.

Comments made by those being interviewed about their use of measurement often reveal most about the nature of the data being collected and used.

Where measurement is dysfunctional what constitutes 'good' or acceptable data and 'bad' or unacceptable data will be clear, and data will conform to 'good'. In extreme cases of dysfunction data is collected, formatted and reported upwards through the organization with little use, the appearance of data being delivered and passed on being more important than its accuracy, completeness or use, care being taken by data collectors to report, as closely as possible to expected norms. In these environments decision making and management control is undertaken using other, less abstract, but also less insightful means.

The following supplementary questions can provide indicators of the perceived value of the data.

1. What are the <data type> definitions
2. How accurate does the data need to be?
3. How accurate is it?

These questions provide some indication of the ownership of the data and investment in its collection and use. If the interviewee knows well what the definition is, or knows where to find out quickly this indicates a degree of involvement with the data. Similarly an understanding of required accuracy and the need for this (as distinct from reporting precision) may indicate that the data has value.

After speaking to various roles throughout an organization a picture of data collection, use, value and 'transport', signalling areas of high and low value, and areas of dysfunction will emerge.

The author has noted that the character of measurement within an organization is often unrepresentative of the character of the organization. Organizations with a positive and exemplary culture can often have dysfunctional measurement simply because it has unrealistic expectations of measurement and even if it recognizes dysfunction does not know how to deal with it, but perseveres with a burdensome system.

4 Eliminating Dysfunction

The recognition of dysfunction (and value), is an essential first step. While requiring some skill, is relatively straightforward. Eliminating it, or avoiding it requires a fundamental change of attitude that can take more time to achieve.

There are three elements to the elimination of dysfunction:

1. Establish a climate or culture for good measurement
2. Support technically correct measurement
3. Evaluate measurement policies and practice periodically – reinforce and educate as required

A culture conducive to good measurement can be established using a number of mechanisms. Establishing 'policies' for good measurement may be used by those familiar with CMMI. The MA PA requires this and policies that encapsulate the principles of informational measurement, data privacy, aggregation and technical adequacy can be developed. But be warned; this may take time, requiring education and persuasion. Many, including measurement experts, will be very uncomfortable with this. It runs counter to our notions of openness, visibility and transparency.

An example of , measurement policies developed by an organization using CMMI are shown below

1. All data will be unambiguously specified to meet the measurement needs. These specifications will be available and accessible to those collecting, analysing and using the measurement data.

2. All collected, analysed and stored data will be traceable back to measurement goals, which are, in turn, traceable to needs

3. Data and any resultant measures will remain private to the individual, project or department concerned unless they are agreed to be helpful in improving the performance of <org> as a whole

4. Measures will be routinely reviewed to determine their benefit. Where the cost of collection outweighs the benefit collection will stop

The development of measurement policies provides an opportunity to explore good measurement practice but further explanation, justification and education will be required. Management in particular may require convincing of the need to acknowledge the information horizon.

Technical correctness has an unexpected role to play in establishing good measurement. In particular the GQM measurement method - often implicated in measurement dysfunction when applied as the basis of a systemic organizational measurement programmes – can be used as a framework and toolset for developing functional measurement (Ref 5). It helps put in place clear processes, techniques and infrastructure for the development of good informational measurement.

If effective policies (or other means for promoting good measurement practice), together with appropriate practices can be introduced then these need to be reinforced. Periodic reviews of measurement practice and the ability to recognize re-emerging dysfunction or challenges to good practice should be undertaken periodically or when measurement problems manifest. Be prepared to take remedial action – education, promotion and promulgation as needed. Constant vigilance is necessary if good measurement practice is not to be lost.

5 Closing Remarks

Measurement practice is capable of major improvement. There is too much unintentional misuse. This misuse is widely felt and discourages others from using measurement. We have failed to heed the warning of Gerry Weinberg:

'If Software Metrics leads to something like Taylorian "scientific management" in software development, I for one, will bow my head in shame.'

Gerald M. Weinberg
writing in the foreword of Toms Gilb's 'Software Metrics', the first s/w metrics book

We need to do something to rectify this and enable the effective use of this valuable tool. I believe we now understand one of the fundamental problems with measurement, and can now begin to fix it.

All organizations can undertake the simple first step and investigate their measurement practice to reveal dysfunction if it exists.

It is then up to those few organizations willing to take the next step and eliminate poor measurement practice to show the rest what measurement can do in helping organizations achieve their goals....

'The good news is that you can succeed in producing a culture conducive to measurement. There are organizations in which people have given themselves completely to the pursuit of organizational goals... ... organizations in which members hunger for measurement as a tool that helps get the job done... ... To use measurement inappropriately would betray a sacred trust, and no one would consider such a betrayal.'

Robert D. Austin

6 References

1. Measuring and Managing Performance in Organizations by Robert D. Austin, pub. Dorset House Publishing, ISBN 0-932633-36-6
2. On the Folly of Rewarding A, While Hoping for B by Steven Kerr, Academy of Management Journal, Vol 18, Number 4, pp769-783
3. Out of the Crisis by W. Edwards Deming, pub Cambridge University Press, 1982, ISBN 0-521-30553-5
4. I Want You To Cheat by John Seddon, pub Vanguard Press, 1992 ISBN 0-9519731-0-X
5. OSEL GQM/Measurement Tutorial, 2007.