Determinants of Success in Software Measurement Programs: Initial Results

Dennis R. Goldenson
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA
dg@sei.cmu.edu

Anandasivam Gopal
Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA
anandg+@andrew.cmu.edu

Tridas Mukhopadhyay
Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA
tridas+@andrew.cmu.edu

Abstract

While a great deal is known about technical issues of data gathering and applied statistics, less is known about what it takes to implement a successful software measurement program. Indeed a good deal of anecdotal evidence suggests that such efforts often fail. In this paper we report the initial results from a large scale survey of practitioners and users of software measurement programs. A preliminary multivariate analysis examines differences in the use of software measurement results in organizational decision making. Three variables account for two thirds of the observed variance.

The research problem

Particularly in these days of diminishing resources and outsourcing, an active program of measurement and analysis is often regarded as critical to the success of software development, maintenance, and acquisition efforts. This is so not just for large-scale, complex, mission critical systems, but for the success of commercial software enterprises as well.

In fact, a great deal is known about technical issues of data gathering and applied statistics as they are or ought to be applied to software measurement and analysis. Well grounded guidance about what constitutes best practice in this area dates back to the development of the statistical sciences in the last century, with much older philosophical underpinnings.

However less is known about what it takes to implement a successful software measurement program. Indeed a good deal of anecdotal evidence suggests that such efforts often fail [12].

Of course useful expert guidance about how best to implement a software measurement effort exists in the form of case studies, summary experience reports, and from a few systematic empirical studies. However the experts too often disagree and a great deal more remains to be learned.

Precious little empirical evaluation and defensible data are available in a field that prides itself on the importance of measurement. There is a need for wider, more rigorous empirical test to provide more confidence in our assertions and guidance to practitioners

We have conducted a broad based survey of practitioners and users of software measurement programs. The sample of 228 includes representatives of defense and other government organizations, defense contractors, and commercial enterprises.

Characterizing the success of software measurement programs

By success we mean more than longevity and persistence over time. To what extent are measurement and analysis regularly used to inform management and technical decision making? Technically defensible shelfware is not enough. Moreover, to what extent can improvements in an organization's performance (e.g., defect density, cycle time, accuracy in forecasting budget and schedule, or operational availability) be attributed to the use of measurement and analysis in that organization?[1]

As can be seen in Table 1, we created two composite indices that summarize six and ten related survey items respectively. These and all subsequent composite variables are simple weighted averages based on the cardinal values of their component items. Principal components analyses indicate that both sets of items are in fact internally consistent in this sample.

Notice also in Figure 1 that there is reasonably wide variation in both composite variables. We can be reasonably confident on face validity grounds that the respondents are answering candidly when they characterize lack of success in their measurement programs. Moreover, the variation leaves room for meaningful statistical analyses of differences in reported success.

Our focus in this paper is on the first of the two composite success variables, namely on the use of measurement and analysis in informing management and technical decision making. As can be seen in Figure 2, there is in fact a reasonably strong relationship between use in organizational decision making and the subsequent improvement of organizational performance ($r^2 = .46$, $p < .0001$). However our explanatory variables are more proximate in time and theoretically to varying use of measurement than they are to subsequent impact on organizational performance.[2]

Explaining differences in program success

The larger study of which this paper is a part examines a wide range of possible explanatory variables. Here we limit our concern to three sets: (1) alignment of the measurement program with wider business and organizational goals; (2) organizational commitment and resource sufficiency; and (3) the technical characteristics of the measurement program itself.

Alignment with business goals

The importance of alignment of the measurement program with the business and technical goals of the organization is a fundamental tenet of software measurement practitioners [1, 2, 5, 10]. We asked our respondents a series of questions about the involvement of various potential stakeholders in setting goals and deciding on plans of action for software measurement and analysis in their organizations.

As can be seen in Table 2, four items ask about the involvement in such agenda setting of the intended users of the measurement results. Two other items

[1] Clearly the inter-relationships are complex and it is unreasonable to expect anything approaching to a one-to-one relationship, but at least some demonstrable impact on business value is needed to justify continued investment.

[2] At this writing, we are currently beginning a more complete analysis of the impact of software measurement results on subsequent organizational performance.

ask about the providers of those results, namely measurement specialists and people from other technical support units. Principal components analysis confirmed our expectation that we should calculate two separate composite indices here.

Both measures are in fact related to successful use of measurement results in informing management and technical decision making. Not surprisingly given GQM and related theory, involvement of the intended users is more strongly related ($r^2 = .42$, $p < .0001$) than is involvement of the providers ($r^2 = .21$, $p < .0001$).

One might argue that we need collaborative involvement by both groups in order to achieve success. However including both variables in a multivariate analysis does not improve our ability to account for differences in actual use of the measurement results. Neither did we find any apparent interaction effects with our simple MANOVA models.[3]

We were also concerned that involvement in setting the agenda may sometimes be counter productive. Hence we asked a similar series of questions about which of the these same groups of potential stakeholders were a source of conflict. At least in this particular sample, there was relatively little evidence of such contention ($r^2 = .01$, $p = .23$). Once again, a multivariate analysis of variance found no apparent interaction effects and no improvement in our ability to account for differences in reported use of the software measurement results.

Organizational commitment and resource sufficiency

The importance of management commitment and the existence of sufficient organizational resources are commonly emphasized as being crucial for software process improvement [4, 6]. Their role in the success of software measurement efforts would appear to be no less important.

Table 3 summarizes the wording of the two items that we included in a composite index of commitment demonstrated by management to their organizations' software measurement efforts. As can be seen there, we also created a four item index of the degree of cooperation and support that was forthcoming from technical people in those same organizations.

As expected, the measure of management commitment is in fact rather strongly related to use of software measurement results ($r^2 = .47$, $p < .0001$). There are also moderately strong bivariate relationships between use of software measurement results and available funding[4] ($r^2 = .20$, $p < .0001$), the quality of measurement related training[5] ($r^2 = .18$, $p < .0001$), and

[3] We ran this and subsequent ANOVA's treating the effect variables as both continuous and categorized distributions. Categorization added no explanatory power over the single multiplicative interaction terms.

[4] "Is sufficient funding available for measurement and related activities in your software organization?" Response alternatives were "almost always (greater than or equal to 80%)," "frequently (greater than or equal to 60%)," "about half of the time (greater than 40% but less than 60%)," "occasionally (less than or equal to 40%)," and "rarely if ever (less than or equal to 20%)."

[5] "How would you best characterize the measurement related training that is available in your organization?" Response alternatives were "excellent," "good," "adequate," "fair," and "poor."

availability of qualified measurement personnel[6] ($r^2 = .20$, $p < .0001$).

The latter relationships may be weaker because the single item variables are less well distributed and more prone to unreliability than are the composite variables. Regardless, we found no apparent interaction effects, and the single item variables do not contribute to our ability to account for differences in reported use of software measurement results.

In addition, the bivariate relationship between the technical support measure and actual use as reported by our respondents is quite low ($r^2 = .10$, $p < .0001$) and it does not contribute to our multivariate results. Our conjecture for now is that the degree of cooperation and support forthcoming from the technical people is more important in explaining the integrity of the data collected than in accounting for variations in use of the software measurement results.

We looked at one final single item variable in this context. The importance of having a well respected measurement "guru" who also understands the organization's business sector and domain is often cited as being crucial for explaining the success of software measurement programs [7, 9, 11]. Perhaps surprisingly, however, the presence of such an individual as measured here is only weakly related to reported use of measurement results in management and technical decision making[7] ($r^2 = .07$, $p < .0001$). Once again, we conjecture that the observed relationship may be attenuated by the way we worded this single question and its related measurement unreliability. However it may also be that the presence of such in-house expertise is simply too uncommon to contribute to a general explanation.

Technical characteristics of the measurement program

Finally, for this initial analysis, we examined a series of technical characteristics of the measurement program itself. As seen in Table 4, these include a five item composite index of the use of a variety of data analytic methods,[8] a three item index of reliance on automated support of the organizations' software measurement activities, a two item index of the extent to which the organizations' data gathering procedures are well defined, and a two item index of quality of the software measurement data.

---

[6] "Are qualified, well-prepared people available to work on software measurement in your organization (i.e., people with sufficient measurement related knowledge, competence, and statistical sophistication)?" Response alternatives were "almost always (greater than or equal to 80%)," "frequently (greater than or equal to 60%)," "about half of the time (greater than 40% but less than 60%)," "occasionally (less than or equal to 40%)," and "rarely if ever (less than or equal to 20%)."

[7] "Measurement has been championed by a well respected "guru" (or gurus) who also knows the organization and it's business." Response alternatives were "almost always," "to a large extent," "to some extent," "to a limited extent," and "hardly at all."

[8] The fourth and fifth component items arguably does not belong in the same common factor as do the other items on which we based this composite index. In fact, they load more heavily (0.92 and 0.77 respectively) on a second factor in our confirmatory analysis. For now, we use a single composite index on predictive validity grounds, since it is in fact strongly related to the criterion variable. Moreover, our conjecture is that all five items will in fact prove to fit into a unidimensional construct as part of an upcoming cumulative (Guttman) scale analysis.

All four variables are in fact related to our criterion measure of successful use. The extent of use of varying data analytic methods is in fact the most strongly related of the four ($r^2 = .48$, $p < .0001$).

The other bivariate relationships are also reasonably strong for data of this kind: reliance on automated support ($r^2 = .21$, $p < .0001$), well defined data gathering procedures ($r^2 = .33$, $p < .0001$), and data quality ($r^2 = .28$, $p < .0001$). However, at least partially due to multicolinearity, they contribute very little as a group to our overall ability to account for actual use of software measurement results in organizational decision making.

Putting it all together: An initial multivariate analysis

Recall that three predictor variables are most strongly related to our proximate criterion of success in implementing software measurement programs. As seen in Figure 3, these are (1) user stakeholder involvement in setting the organization's measurement agenda, (2) management commitment, and (3) the extent of use of varying data analytic methods.

Based on these bivariate results and preliminary multivariate analyses, we settled on a single, simple MANOVA to summarize variation in reported use of software measurement in our respondents' organizations. As seen in Figure 4, the model includes only three predictor variables, one from each of the three sets of potential predictors we initially considered.

The main effects of these three variables account for almost two thirds of the observed variance in our criterion index of use of software measurement results. There is some multicolinearity among the three predictors, but the variance explained is noticeably higher than is so for any of the single strongest bivariate relationships.

We are unable to identify any significant interaction effects. Moreover, adding other main effects into a more complex model adds essentially no improvement in overall explanatory power ($R^2 = 68$).[9]

Conclusions and next steps

Based on our preliminary analysis, we have identified a simple multivariate model that is capable of explaining variation in our criterion index in a very parsimonious manner. The main effects of three variables account for two thirds of the observed variance in our index of differences in the use of software measurement results in organizational decision making.

While these results may seem to be fairly intuitive, we have added a better quantitative description than was available previously. Moreover we have failed to find evidence to support other commonly stated assertions about what it takes to establish a successful software measurement program.

Of course, much more remains to be done. One most certainly ought not to conclude that only three variables are all that matter. Much more is needed for a

[9] Four additional variables had statistically significant main effects in preliminary MANOVA's limited to each of the initial three sets of potential predictors. These are (1) stakeholder involvement by measurement providers in setting the organization's software measurement agenda , (2) availability of qualified measurement personnel, (3) well defined data gathering procedures, and (4) quality of the software measurement data. Only provider stakeholder involvement remains significant in the final preliminary MANOVA we examined.

fuller explanation of what it takes to establish a successful software measurement program.

We know a great deal about the technical attributes that constitute best practice in software measurement. But we need a much better understanding of what it takes to generate adequate management support and stakeholder participation in setting the measurement agenda. For that matter, we need to know more about what it takes to build technically competent measurement programs.

Methodologically, we continue to conjecture that more robust data analytic methods better suited to categoric variables will in fact help us identify important interactive and precedence relationships, and thus improve our ability to provide better guidance to the practitioner community.

Our current research agenda includes simultaneous and structural modeling based on contingency and institutional theory perspectives from management science [3, 8], and a cumulative scale analysis of CMM® related assertions about appropriate precedence in doing software measurement.

Moreover we have a great deal of survey data yet to analyze. Pertinent areas include measurement's role in software process maturity and in larger organizational improvement programs, the impact of requirements imposed by customers and various other forces outside of the organization, the residual effects of earlier failed measurement efforts, "organizational articulation" (i.e., alignment, coordination, and communications patterns), organization size, sector, and domain. We are particularly interested in various institutionalization factors whose effects may be expected to differ in context.

Appendix: About the survey and the sample

The sample is drawn from three overlapping lists of individuals who were thought to be knowledgeable about the measurement programs in their software organizations. The lists were provided in confidence by three sources with wide access to such information.

As seen in Figure 5, the lists include people who represent a wide variety of government and commercial organizations, but they are organizations who were selected from a decidedly atypical subset of the wider software community. Our purpose was to ensure sufficient variation to allow meaningful multivariate analyses. The results have limited generalizability at best in describing the state of the practice in organizations that have only limited experience with software measurement.

The survey was administered via the world wide web from late November 1998 through February of 1999 with a response rate of 55%. The response rate is reasonably high by current standards in survey research, especially given the time of the year when it was done.

Moreover the true response rate is probably somewhat higher. Some of those who failed to reply undoubtedly do not meet our criterion of familiarity with their organizations' measurement programs. In addition, some of the older email addresses may be inactive without proper administrative closure.

References

1. Briand, L.C., Differding, C.M. and Rombach, H.D. 1996. Practical Guidelines for Measurement-Based Process Improvement. *Software Process - Improvement and Practice*. Vol. 2. Pp. 253-280.
2. Daskalantonakis, M. 1992. A Practical View of Software Measurement and Implementation Experiences Within Motorola. *IEEE Transactions on Software Engineering*. Vol. 18, 11 (November).
3. Drazin, R. and Van de Ven, A.H. 1985. Alternative Forms of Fit in Contingency Theory. *Administrative Science Quarterly*. Vol. 30. Pp. 514-539.
4. El-Emam, K., Goldenson, D., McCurley, J., and Herbsleb, J. Success or Failure? Modeling the Likelihood of Software Process Improvement. Under review for publication. February 1999.
5. Hall, T. and Fenton, N. 1997. Implementing Effective Software Metrics Programs. *IEEE Software* (March/April). Pp. 55-65.
6. Herbsleb, J., Zubrow, D., Goldenson, D., Hayes, W., and Paulk, M. 1997. Software Quality and the Capability Maturity Model. *Communications of the ACM* (June).
7. Jeffery, R. and Berry, M. 1993. A Framework for Evaluation and Prediction of Metrics Programs Success. *Proceedings of the First International Software Metrics Symposium*. IEEE Computing Society Press. Los Alamitos, CA.
8. Lawrence, P.R. and Lorsch, J.W. 1967. *Organization and Environment: Managing Differentiation and Integration*. Graduate School of Business Administration, Harvard University, Cambridge, MA.
9. Offen, R.J, and Jeffery, R. 1997. Establishing Software Measurement Programs. *IEEE Software* (March/April). Pp. 45-53.
10. Pfleeger, S. L. 1993. Lessons learned in Building a Corporate Metrics Program. *IEEE Software*. Vol. 10, 3.
11. Rifkin, S. and Cox, C. 1991. Measurement in Practice. *Technical Report-91-16*. Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.
12. Rubin, H. 1993. Debunking Metric Myths. *The American programmer (February)*.

Acknowledgements

First and foremost we offer our thanks to our respondents. Work of this kind simply would not be possible without their candid and willing cooperation. Many others provided much appreciated assistance in a variety of ways, including clarifying the study design and data analysis, construction of the sample, development and online implementation of the questionnaire, and the survey pretest. Thanks are due in particular to Khaled El-Emam, Wolf Goethert, Will Hayes, David Herron, Jay Huber, Carol Jarosz, Cheryl Jones, Mark Kasunic, Jack McGarry, Wayne Middleton, Jim McCurley, Bob McNeill, Raghav Nandyal, Tod Pike, Stan Rifkin, Carrie Ross, Jim Rozum, David White, Dave Zubrow, Michael Zuccher and an anonymous referee.

The Software Engineering Institute (SEI) is a federally funded research and development center sponsored by the U.S. Department of Defense and operated by Carnegie Mellon University.

Tables and Figures

Table 1: Questionnaire items, reliability scores, and factor loadings for study dependent variables

Use in decision making and management		
Cronbach's Alpha = 0.74		
How widely are software measurements actually used in making management and development decisions?		***Factor Loadings***
a)	Monitoring and managing individual projects or similar work efforts	0.76
b)	Use of historical data for project planning and estimation	0.75
c)	Rolled up for larger organization and enterprise wide purposes	0.70
d)	For use by individual engineers, programmers and other practitioners	0.67
e)	Changes are made to technologies, business or development processes as a result of out software measurement efforts	0.72
f)	Staffing and personnel changes are made due to measurement efforts in our organization	0.52
Organizational performance		
Cronbach's Alpha = 0.94		
In your judgment, how much has the use of software measurement improved your organization's performance?		***Factor Loadings***
a)	More accurate budget estimates or ability to reduce costs	0.75
b)	More accurate schedule estimates or ability to reduce cycle time	0.83
c)	Better adherence to customer or user requirements or improved customer satisfaction	0.82
d)	Fewer software defects, faults or failures	0.88
e)	Better functionality or user interface	0.82
f)	Better over-all quality of products and services	0.88
g)	Improved staff productivity or reduced rework	0.84
h)	More informed judgments about the adoption or improvement of work processes and technologies	0.80
i)	Better work processes	0.80
j)	Better strategic decision-making	0.79

Figure 1: Univariate distributions for study dependent variables

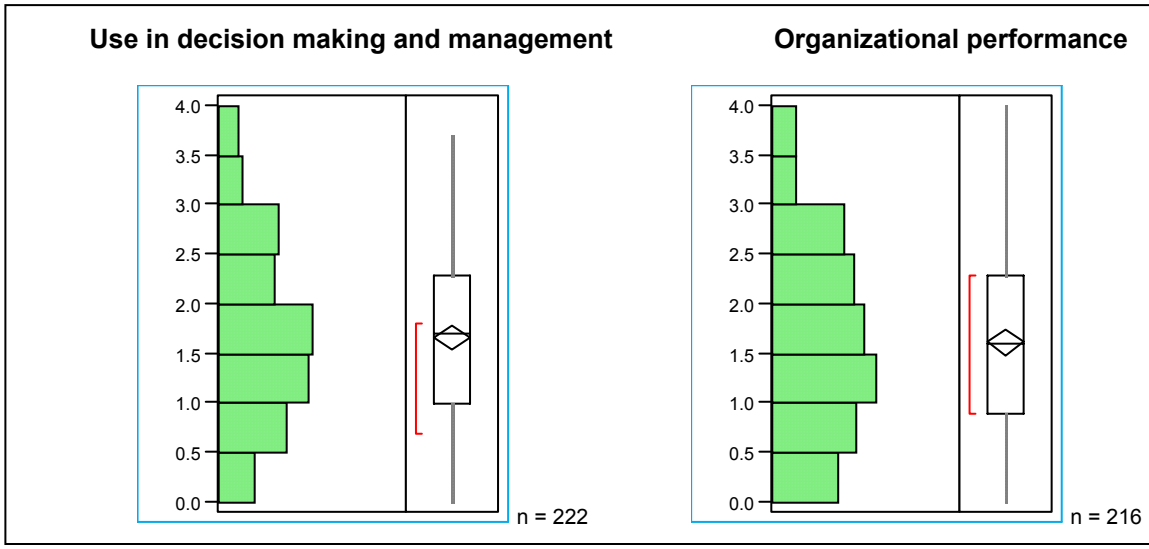


Figure 2: Relation between use of software measurement in decision making and subsequent organizational performance

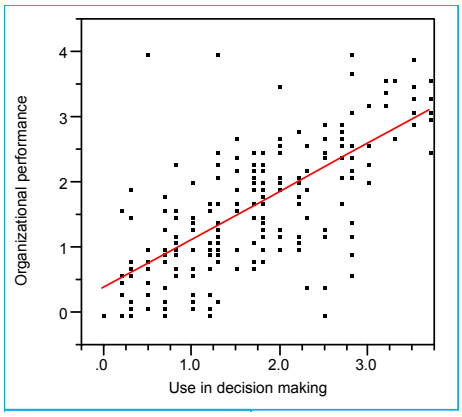


Table 2: Questionnaire items, reliability scores and factor loadings for alignment with business goals

Aligned with intended users		
Cronbach's Alpha = 0.70		
How would you characterize the involvement of various potential stakeholders in setting goals and deciding on plans of action for measurement in your organization?		***Factor Loadings***
a)	Senior enterprise and organization level managers	0.68
b)	Project level managers	0.73
c)	Individual engineers, programmers or other practitioners	0.75
d)	Business support units, e.g. Finance, marketing	0.60
Aligned with measurement and technical people		
Cronbach's Alpha = 0.85		
How would you characterize the involvement of various potential stakeholders in setting goals and deciding on plans of action for measurement in your organization?		***Factor Loadings***
a)	Technical support units	0.84
b)	Measurement specialists	0.84

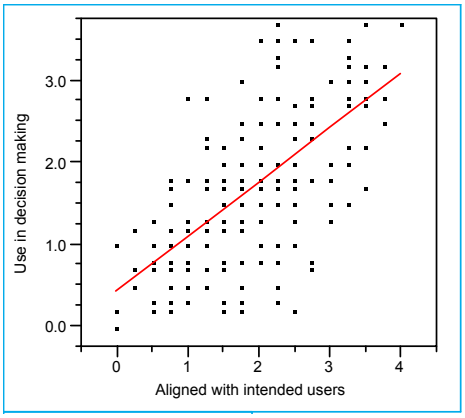
Table 3: Questionnaire items, reliability scores and factor loadings for organizational commitment and resource sufficiency

Management commitment		
Cronbach's Alpha = 0.83		***Factor Loadings***
a)	Management regularly monitors the progress of software measurement activities	0.92
b)	Management clearly demonstrates commitment to measurement	0.93
Technical support		
Cronbach's Alpha = 0.75		***Factor Loadings***
a)	The effort required for people to submit data is often considered to be onerous or burdensome (reverse-scored)	0.79
b)	People consistently provide information as planned and when requested	0.67
c)	The way software measurement data are collected and used is often considered to be inappropriate by the people who must provide the information (reverse-scored)	0.76
d)	There is resistance to doing measurement here (reverse-scored)	0.78

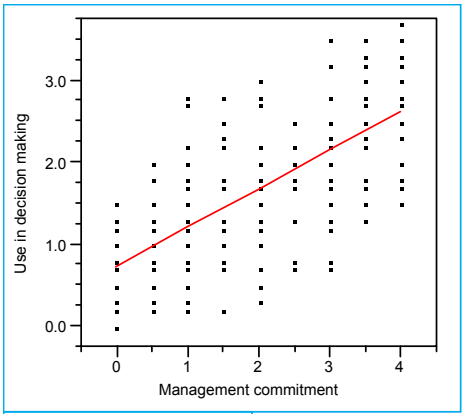
Table 4: Questionnaire items, reliability scores and factor loadings for technical characteristics of the measurement program

Use of analytic methods		
Cronbach's Alpha = 0.76		***Factor Loadings***
a)	Comparisons are regularly made between current project performance and previously established performance baselines and goals	0.67
b)	Sophisticated methods of analyses are used on a regular basis	0.89
c)	Statistical analyses are done to understand the reasons for variations in performance	0.89
d)	Experiments and/or pilot studies are done to prior to widespread deployment of major additions or changes to development processes and technologies	0.04
e)	Evaluations are done during and after full-scale deployments of major new or changed development processes and technologies	0.37
Reliance on support tools		
Cronbach's Alpha = 0.79		
How much automated support is available for software measurement related activities in your organization?		***Factor Loadings***
a)	Data collection (e.g., on-line forms with "tickler" reminders, time stamped activity logs, static or dynamic analyses of call graphs or run-time behavior	0.85
b)	Data management (e.g., distributed database packages, open database connectivity, tools for data integrity, verification, and validation	0.87
c)	Data analysis and report preparation (e.g., spreadsheets, statistical, graphing, and report presentation packages)	0.81
Well defined data gathering procedures		
Cronbach's Alpha = 0.88		***Factor Loadings***
a)	Responsibilities and procedures for recording our measurement data are clearly stated and well understood	0.95
b)	Our procedures for data collection and submittal are well-integrated with other responsibilities and work processes	0.95
Data quality		
Cronbach's Alpha = 0.70		***Factor Loadings***
a)	The data we collect are often inaccurate or incomplete	0.88
b)	Our measures are well defined to accurately capture what they are meant to portray	0.88

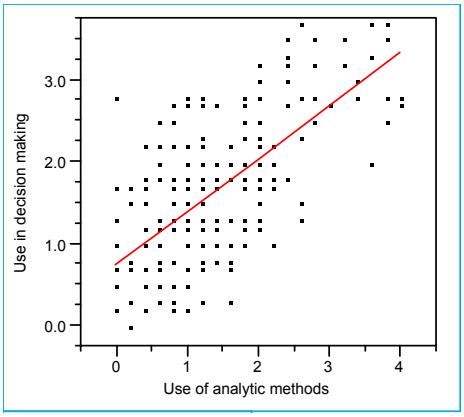
Figure 3: Characteristic bivariate relationships



r^2 = .42, p = .0001, n = 221

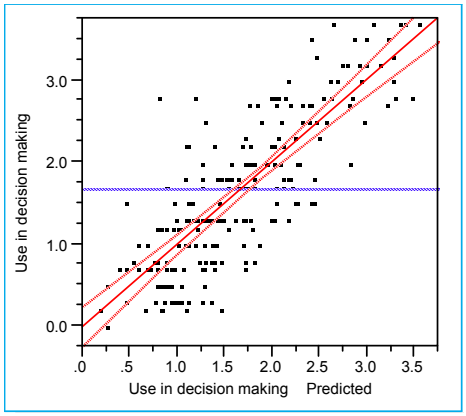


r^2 = .47, p = .0001, n = 221



r^2 = .48, p = .0001, n = 222

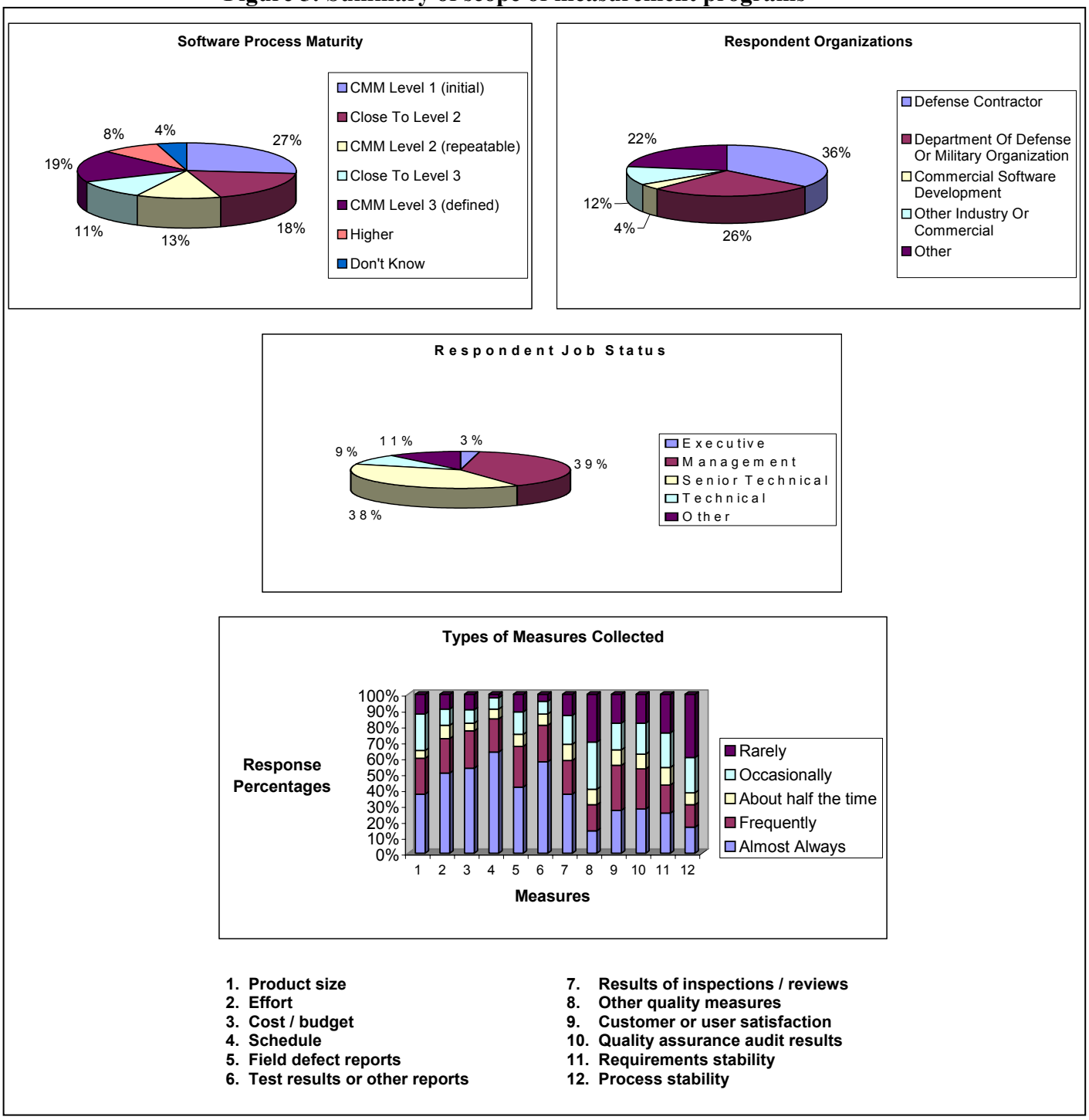
Figure 4: Summary multiple analysis of variance



R^2 = .66, F = 142.25, p < .0001, n = 220

Term	Std beta	Prob > | t |
Aligned with intended users	0.27	< .0001
Management commitment	0.33	< .0001
Use of analytic methods	0.37	< .0001

Figure 5: Summary of scope of measurement programs



1. **Product size**
2. **Effort**
3. **Cost / budget**
4. **Schedule**
5. **Field defect reports**
6. **Test results or other reports**
7. **Results of inspections / reviews**
8. **Other quality measures**
9. **Customer or user satisfaction**
10. **Quality assurance audit results**
11. **Requirements stability**
12. **Process stability**