Putting Schedule Quality Checks to the Test

Eric Lofgren, Technomics, Inc. – ICEAA 2016

Analysts assess schedule quality in order to gauge the reliability of project duration information. Poor quality schedules lack the ability to accurately incorporate new project data, distorting the measure of work-scope progress relative to plan. Further, the forecast of project duration becomes unreliable. Stakeholders are eager to understand the quality of a schedule so that they can gauge whether its assessment of project information is sound enough for decision making.

This study took a sample of 19 defense contract schedules and separated them into two subjective groups, good actors and bad actors. Over the first 10% of the contract, the good actors had a mean duration forecast error of 35%, the value for bad actors was 45%. The baseline forecast error discrepancy between good and bad actors might be significant, but is not especially striking. Yet over the next 10% increment of the contract duration, the good actors quickly incorporated project information and on average reduced the forecast error by 26 percentage points to only 9%. The bad actors, however, scarcely picked up on project information, and their average forecast error only fell by 1 percentage point to 44%. In fact, the bad actors achieved not until 80% the mean forecast accuracy of the good actors at 20%. See Figure 1 below.

Stakeholders would like to know whether their schedule is a good actor, providing confidence that they can base important decisions on accurate and timely information, or a bad actor, which can lead to misinformed decisions. The U.S. Defense Contract Management Agency (DCMA) recommends a 14-Point Assessment for schedule quality that has become the industry standard. This paper will explore how well the 14-Point Assessment predicts a schedule forecast's accuracy and timeliness. It finds that the 14-Point Assessment provides a model with relatively robust parametric statistics, but lays its reliability suspect by applying further tests. The analysis suggests that a larger sample size can return consistent estimates for the 14 coefficients, but the total amount of variation explained will be relatively small because one can never fully separate schedule quality from contract assumptions and performance. Despite the findings, the continued use of heuristics like the 14-Point Assessment will be advocated and expanded to include longitudinal schedule analyses.

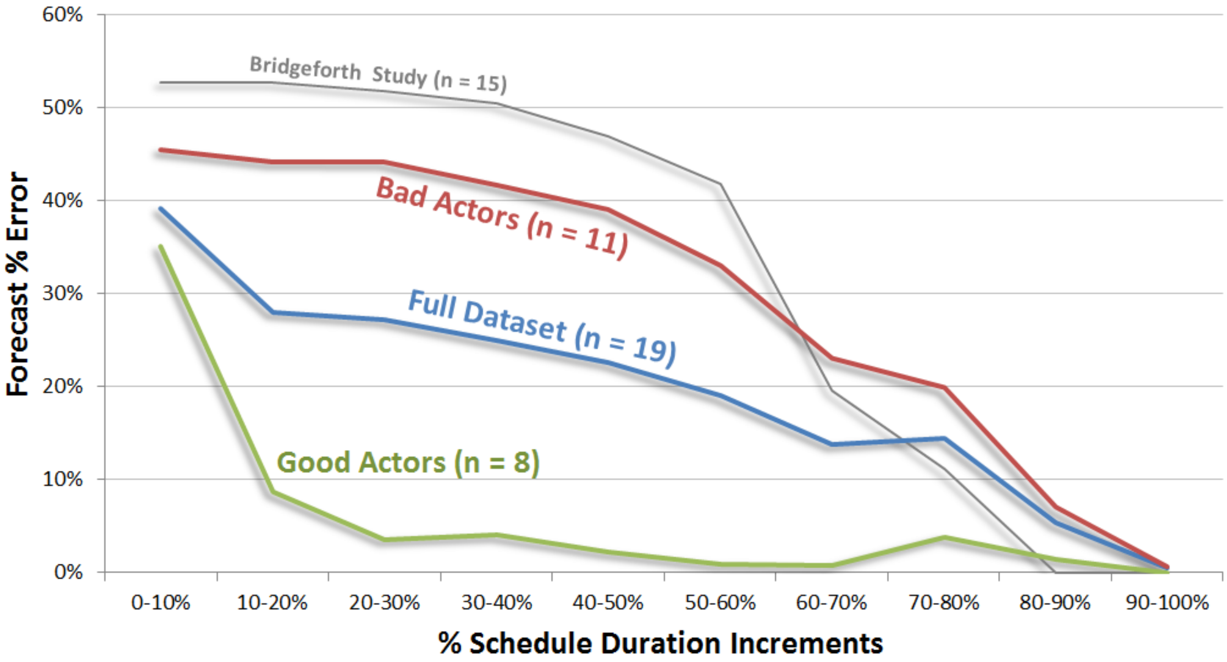


Figure 1: Evolving View of Schedule Forecast Error

Background and Purpose

Contracts for major defense acquisition programs tend not only to be large in value and long in duration, but relatively risky ventures. For stakeholders to make informed decisions, the government requires defense contractors to implement an Earned Value Management System (EVMS) on all large contracts. The foundation of EVMS is the work package, the basic planning unit which represents a functional element, or a well-defined portion of the total contract work-scope. Managers forecast time-phased budgets for each work package, to which actual expenditures are compared. Yet managers also plan to a finer level of detail, breaking the work package down into discrete tasks with relational dependencies, or logic links, in what is called a networked schedule. The completion of schedule activities informs the progress of scope in the work packages. The process is depicted in Figure 2 below. EVMS produces two reports, a Contract Performance Report (CPR) on the progress in the cost of work packages, and an Integrated Master Schedule (IMS) on the progress of timing in networked schedule tasks.

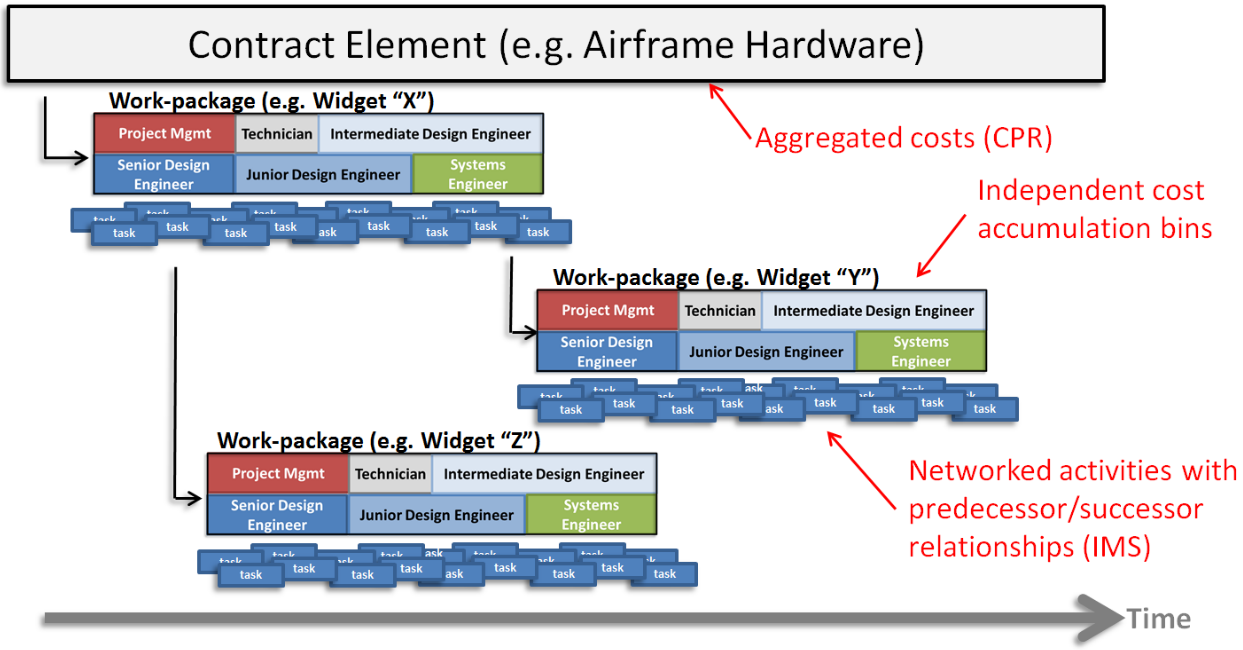

Contract Element (e.g. Airframe Hardware)
Work-package (e.g. Widget "X")
Work-package (e.g. Widget "Y")
Work-package (e.g. Widget "Z")
Project Mgmt
Technician
Intermediate Design Engineer
Senior Design Engineer
Junior Design Engineer
Systems Engineer
task
Aggregated costs (CPR)
Independent cost accumulation bins
Networked activities with predecessor/successor relationships (IMS)
Time

Figure 2: Depiction of the EVMS Process

The two EVMS reports form the basis of contract performance analysis. Because each work package is also tied to a piece of contract scope which represents a functional element, one can use EVMS to gauge the three primary areas of contract risk: cost, technical, and schedule. In general, one can only offset realized risk in any one area at the expense of performance in one or both of the others (see Figure 3 below). For example, if the contractor forecasts a significant schedule delay, the government manager may allocate additional resources to the problem or reduce systems capabilities in order to maintain the contractual delivery date. Therefore, accurate and timely information on schedule risk provides flexibility in project management.

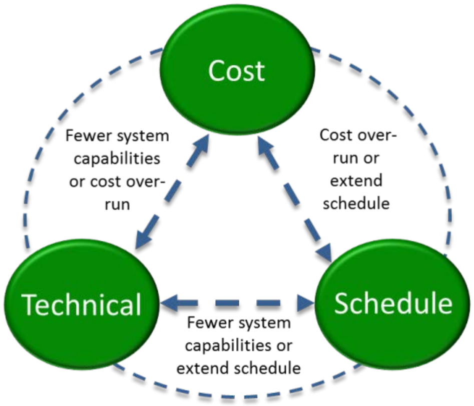

Cost
Technical
Schedule
Fewer system capabilities or cost over-run
Cost over-run or extend schedule
Fewer system capabilities or extend schedule

Figure 3: Cost, Schedule and Technical Risk Interrelationship

Within EVMS, only the IMS can answer the question “when will the deliverable arrive?” As described above, the IMS is a networked schedule where discrete tasks are detailed and provided interrelationships such that the effect of a slip in one task reverberates throughout the schedule and the total impact is calculated. Because the IMS is a living document where only the latest submission incorporates up-to-date information, analysts tend to discard all past submissions. For this reason, authoritative schedule guidance has focused on testing the quality of the current submission.

All major defense contracts go through an Integrated Baseline Review where the government scrutinizes the contractor’s cost and schedule reports. The 14-Point Assessment tends to be the framework on which IMS analysis is based, and might be the first and last check on an IMS. The assessment measures characteristics such as how many tasks are disconnected them from the network, and the proportion of tasks which have finished late relative to their baseline end date. See the Appendix for a full description of each of the 14 checks.

Even with the 14-Point Assessment, analysts on both the government and contractor side find it difficult to ascertain the realism of any given schedule. Though threshold values are provided for each check that may signal a problem in that area, no one knows what the potential impacts to the schedule may be. It has not been tested how well the 14-Point Assessment predicts the ultimate quality of the schedule.

In fact, discourse on schedule forecasting has focused on the cost-based CPR rather than the IMS. The predominance of CPR analysis over IMS has occurred in the defense industry for a few reasons. First, CPRs are used to in cost analysis and help formulate budget requirements which tend to be revised annually, making short-term cost information more valuable than long-term schedule information (even though the latter may ultimately drive the former). Because stakeholders tend to focus on budgets, resources have been put into both auditing CPRs and making the CPRs readily available by developing a query-ready relational database. The IMS as of 2015 is not subject to official scrutiny or auditing, nor is there a relational database from which to retrieve its data. Second, IMS data are far richer than CPR data. While the lowest level of insight in the CPR is an aggregation of work package information, the IMS provides insights into the activities which combine to form a work package. The resolution available through the IMS is therefore an order of magnitude greater than available in the CPR. Further, each activity in the IMS has potentially hundreds of attributes,[1] whereas basic data elements in the CPR only have five.[2]

Countless studies have been published on schedule forecasting techniques and estimating correlates of schedule slip. A great many of these studies, however, have been performed without use of IMS data. They often use contract attributes as independent variables to estimate total contract slip. However, in determining schedule quality, one doesn't only want to measure how poor the first schedule estimate was, but how quickly the schedule incorporated information to provide an accurate estimate.

In "Using Earned Value Data to Forecast the Duration of Department of Defense Space Acquisition Programs," Shedrick Bridgeforth analyzed all monthly IMS submissions for 15 contracts. His primary focus was comparing different EVMS schedule forecast techniques, mostly comprising CPR-based methods, but also one IMS-

[1] Hundreds of attributes if only because each activity relation, or logic link, must be enumerated. Further, each relation has a type, such as Finish-to-Start, Start-to-Start, etc. Others include numerous attributes for each of time-phasing, constraints, float, and resourcing.

[2] The five CPR metrics are actually four: Budgeted Cost of Work Scheduled (BCWS); Budgeted Cost of Work Performed (BCWP); Actual Cost of Work Performed (ACWP); and Estimate to Complete (ETC). The sum of time-phased BCWS equals the Budget at Complete (BAC). In all fairness, the work package itself can have numerous other attributes, such as those that define its cost collection point and resource charges. The current CPR requirements aggregate away much of that detail.

based.[3] In Figure 1 above, the time-phased forecast error from his set of contract IMSs is compared to the current set. Bridgeforth's data set only included space systems, traditionally a high cost/schedule growth area, while the set for this paper spans several commodity groups.

Instead of testing which manipulations to EVMS data can provide the most accurate and timely schedule estimate, as Bridgeforth has done, this paper will attempt to discover whether the 14-Point Assessment correlates with an IMS's realized quality. Specifically, the association between the 14-Point Assessment on the first IMS submission and the IMS's accuracy and timeliness will be explored.

Data and Methods

All IMS contract data utilized were accessed through the Earned Value Management Central Repository (EVM-CR), which is resident in the Cost Assessment Data Enterprise (CADE). In order to measure the realized schedule quality, the only usable contracts were those that consistently provided IMS data throughout. There must have been a first submission (usually 60 days after contract award, but often variable), regular submissions throughout, and a final submission (EVMS reporting is not mandatory after 90% of work-scope is completed). All indefinite delivery-indefinite quantity (IDIQ) contracts were neglected because they have task orders put on contract which both changes the scope and duration, but which cannot be separated out from the original scope in the IMS. Ultimately, 19 completed contracts were analyzed consisting of 266 individual IMS submissions.

As stated before, though a relational database of CPR data has been developed, there currently is not one for IMS data. IMS data are submitted by the contractor in their native format (i.e., in the file type in which the schedule was built in). The primary scheduling software that contractors utilize are Microsoft Project, Primavera, and Open Plan. A third-party software called Acumen was used as the common method for extracting IMS data from its native format into a common format. Each IMS extract

[3] The author of this paper developed the one IMS-based method used in Bridgeforth's study, see "Trust, but Verify: An Improved Estimating Technique Using the Integrated Master Schedule" from ICEAA 2014.

included 108 attributes for each and every schedule task. The total set of 266 was then organized into a Microsoft Access database. This data table will be referred to as the "Primary" data table.

With only 19 contract observations (four more than Bridgeforth's study), it is unlikely that a parametric analysis can determine coefficients for 14 independent variables with any degree of certainty. In an attempt to corroborate and strengthen any results, the 14-Point Assessment will be tested against different levels of projects within the IMS. The three project levels are: 1) total project level; 2) subproject level; and 3) task level. The process for setting up the datasets at each level will be discussed in turn.

1. Project Level

The set of independent variables for each project were the output from the 14-Point Assessment on the first IMS submission. The first IMS was used to derive the independent variables because the desire was to see how well the first assessment score predicts the realized accuracy and timeliness of the schedule. The dependent variable for each project, the measure of accuracy and timeliness also used by Bridgeforth, is the Mean Absolute Percent Error (MAPE). Calculating the MAPE requires a longitudinal analysis of IMSs. Each IMS submission's place within the total project duration was calculated in the following way:

$$\%\ of\ Schedule = \frac{(Current\ IMS\ Time\ Now - First\ IMS\ Time\ Now)}{(Actual\ End\ Date - First\ IMS\ Time\ Now)}$$

Total schedule duration, the denominator of the equation above, was calculated as the time between the first IMS submission and the actual end date, rather than the actual start to end date, because the strength of schedule networking is dependent upon its current state. Further, there was a variable lag between the actual start and first IMS submission between contracts. Because the IMS is primarily a forecasting tool, the most reasonable measure of duration is from time of initial forecast to actual contract delivery.

Each IMS submission also provides an estimate of what the project end date will be, which can be translated into a percent error at each point in time. The percent error for each IMS submission was calculated in the following way:

$$\%\,Error = \frac{(Actual\ End\ Date - Current\ IMS\ Forecast\ End\ Date)}{(Actual\ End\ Date - First\ IMS\ Time\ Now)}$$

Because the distribution and density of IMS submissions collected for each contract was slightly different, the mean of the percent errors was not used to calculate the MAPE. Instead, for each contract, the IMS submissions were bucketed into 10% increments and the mean percent error for each increment was calculated. The data collection was tailored such that each contract had at least one observation per 10% increment. The overall contract MAPE was then calculated as the mean of the increments' mean percent error.

The MAPE does a good job of measuring accuracy and timeliness because the quicker an IMS reports schedule slip, the lower the MAPE will be. For example, consider two contracts which both realized a 50% slip. The first contract IMS integrated project information quickly and signaled a 50% slip early, while the other poorly integrated project information forecasted no slip until late. Though the error represented in the first IMS is exactly the same for both, the former IMS will have a relatively low MAPE as it integrated information on potential slip quickly and the latter will have a relatively high MAPE.

2. Subproject Level

Analysis below the total project level is extremely difficult because individual tasks from the first IMS often get dropped over time, or change identifiers in some untraceable way. Therefore, a MAPE can only be calculated for a subproject where the last task to finish appears in either: a) the last schedule; or b) some intermediate schedule where the actual task finish precedes the IMS "time now" date. All tasks which satisfied one or both of the criteria would be the fundamental reference point for a subproject. The rest of the associated tasks in the subproject are the predecessors of that task, which was able to be traced through IMS submissions, as well as the predecessors of that task's predecessors. Note that this immediately introduces a bias as good performing tasks are less likely to be dropped or manipulated than poor performing tasks. Yet this critique also applies at the project level, as poor performing contracts often fail to submit regular IMSs or are cancelled.

To find a task's predecessors, a second extract from Acumen was required which listed all the successor activities for each and every IMS task. To find the immediate first set of predecessors, the successor relationships were simply reversed to return predecessor relationships. Next, the second set of predecessors, or all the predecessors of a task's predecessors, had to be found and tagged to the primary subproject task. Note that individual tasks will often find themselves in multiple subprojects. These subproject relationships from each contract were then joined together and made into the second major data table in the Microsoft Access database. This data table will be referred to as the "Relationships" data table.

The next step was to calculate the 14-Point Assessment for each and every subproject. This required joining information from both the Primary and Relationships data tables. The Primary table includes all the attributes necessary to conduct the 14-Point Assessment while the Relationships table provides the information that determines what tasks enter a subproject. Both data tables were imported into R, and an additional field was added to the Primary table which references the Relationships table. All tasks in a given subproject would have a value of "1" in the new field and those which were not had a value of "0". An iterative algorithm was then developed which calculated the 14-Points for one subproject, stored those results, and moved on to the next until all 13,936 subprojects were completed.

The calculations for the 14-Point Assessment at the subproject level were exactly the same as those at the total project level except for one case. The Critical Path Test checks the integrity of the overall network logic by slipping tasks and observing if there is a proportional slip in the end date. Because subprojects were discovered by using the relationships which the IMS provided, they cannot have broken logic links and thus they cannot fail the Critical Path Test.

The subproject level MAPE was also calculated in the same manner as the total project level. It was already found which tasks had their finish dates realized in subsequent IMS submissions, and these formed the reference point for determining subprojects by finding two sets of predecessors. These tasks, each of which represents the final task in a subproject, are used to calculate the MAPE. Again, the two values needed from each IMS submission are the percent of schedule and the percent error. They are calculated in the exact same way as at the project level, except the end date from the subproject is used instead of the end date for the entire contract. The one major difference is that there was no assurance that there was at least one IMS observation per 10% increment for each and every subproject. For example, if a subproject were to finish in 8 months after the first IMS submission, there wouldn't be enough potential observations to calculate a percent error for each 10% increment.

3. Task Level

Task level calculations were relatively straight-forward, but, because this level represents the performance of an individual task instead of a networked set of tasks, it is the least similar to the total project level. The 14-Point Assessment had to be substantially modified to be applied to individual tasks. Often, variables which are proportions in the project or subproject level were turned into binary variables at the task level. For example, the logic test returns the percent of incomplete tasks missing either a predecessor or successor. At the task level, this would be a zero if the task is missing logic and a one if it has logic. Further, 3 of the 14 variables could not be performed at the task level because they fundamentally require multiple tasks to perform. These are: the Critical Path Test; the Critical Path Length Index (CPLI); and the Baseline Execution Index (BEI). The Appendix includes all proxy calculations performed for the 14-Point Assessment at the subproject and task level.

The MAPE for the task level was calculated in the same way as the project and subproject levels. Again, because tasks often finish during the execution of the contract, there will not always be a percent error observation for each of the 10% increments. Similarly, only tasks which appeared in the final IMS, or had their actual finish date precede an interim IMS, were used in this analysis. A total of 14,332 task level observations were found.

Results

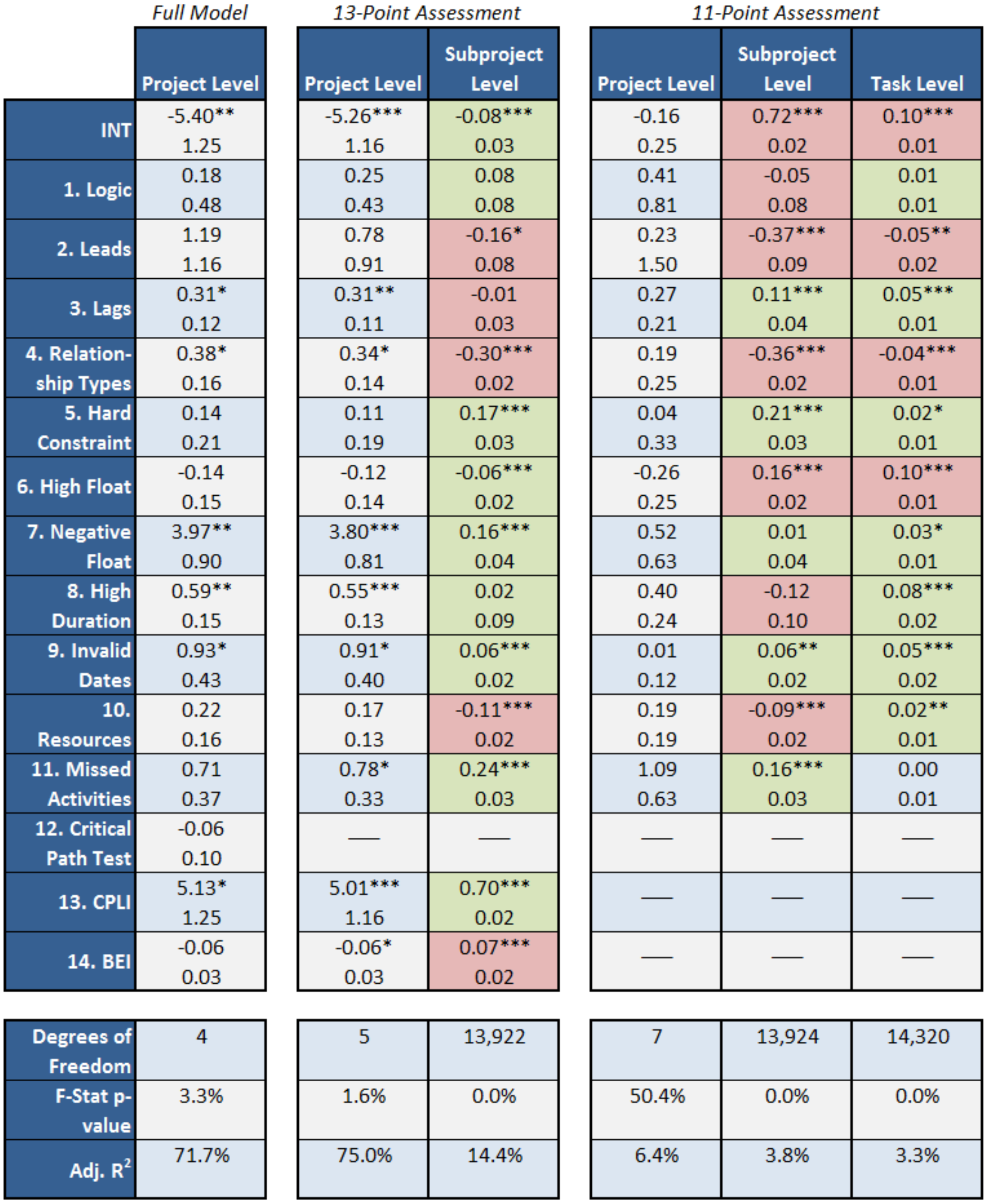
Because there is an extremely large number of possible independent variables that may be tested against the MAPE, from both the IMS and from other contract information, it was desired that the parametric specifications be as simple as possible. Two guiding principles directed the subsequent analysis: 1) no additional control variables beyond the 14 criteria would be tested, especially considering that there are only four degrees of freedom (19-14-1) at the project level; and 2) the data would only be tested against a linear ordinary least squares (OLS) model as a first order best approximation. The results of the regressions at each level are shown in Table 1 below. There were three specifications:

Full Model: includes all 14 variables from the Assessment and was performed on the project level dataset only;

13-Point Check: the full model less variable 12. Critical Path Test which was performed on both the project and subproject level; and

11-Point Check: the full model less variables 12. Critical Path Test, 13. CPLI, and 14. BEI which was performed on all three levels. See the Appendix for a description of which checks cannot be performed at which levels.

The sign of the project level coefficients were stable across all three specifications, though their values differed in some places significantly. Between the Full Model and the 13-Point Assessment, the p-value decreased from 3.3% to 1.6% and adjusted R^2 increased from 71.7% to 75%. Removing the Critical Path Test increased the overall fit of the data. Further, two additional coefficients reached significance at the 90% confidence level and the confidence level for five other coefficients increased to either 95% or 99%. This indicates that multicollinearity between regressors may have caused poor coefficient estimates. The 11-Point Assessment specification, which removed the CPLI and BEI, fit the data poorly, with an adjusted R^2 of 6.4% and no coefficients estimated to be significant at the 90% confidence level.



	Full Model	*13-Point Assessment*		*11-Point Assessment*		
	Project Level	Project Level	Subproject Level	Project Level	Subproject Level	Task Level
INT	-5.40**	-5.26***	🟩 -0.08***	-0.16	🟥 0.72***	🟥 0.10***
	1.25	1.16	0.03	0.25	0.02	0.01
1. Logic	0.18	0.25	🟩 0.08	0.41	🟥 -0.05	🟩 0.01
	0.48	0.43	0.08	0.81	0.08	0.01
2. Leads	1.19	0.78	🟥 -0.16*	0.23	🟥 -0.37***	🟥 -0.05**
	1.16	0.91	0.08	1.50	0.09	0.02
3. Lags	0.31*	0.31**	🟥 -0.01	0.27	🟩 0.11***	🟩 0.05***
	0.12	0.11	0.03	0.21	0.04	0.01
4. Relationship Types	0.38*	0.34*	🟥 -0.30***	0.19	🟥 -0.36***	🟥 -0.04***
	0.16	0.14	0.02	0.25	0.02	0.01
5. Hard Constraint	0.14	0.11	🟩 0.17***	0.04	🟩 0.21***	🟩 0.02*
	0.21	0.19	0.03	0.33	0.03	0.01
6. High Float	-0.14	-0.12	🟩 -0.06***	-0.26	🟥 0.16***	🟥 0.10***
	0.15	0.14	0.02	0.25	0.02	0.01
7. Negative Float	3.97**	3.80***	🟩 0.16***	0.52	🟩 0.01	🟩 0.03*
	0.90	0.81	0.04	0.63	0.04	0.01
8. High Duration	0.59**	0.55***	🟩 0.02	0.40	🟥 -0.12	🟩 0.08***
	0.15	0.13	0.09	0.24	0.10	0.02
9. Invalid Dates	0.93*	0.91*	🟩 0.06***	0.01	🟩 0.06**	🟩 0.05***
	0.43	0.40	0.02	0.12	0.02	0.02
10. Resources	0.22	0.17	🟥 -0.11***	0.19	🟥 -0.09***	🟩 0.02**
	0.16	0.13	0.02	0.19	0.02	0.01
11. Missed Activities	0.71	0.78*	🟩 0.24***	1.09	🟩 0.16***	0.00
	0.37	0.33	0.03	0.63	0.03	0.01
12. Critical Path Test	-0.06	—	—	—	—	—
	0.10					
13. CPLI	5.13*	5.01***	🟩 0.70***	—	—	—
	1.25	1.16	0.02			
14. BEI	-0.06	-0.06*	🟥 0.07***	—	—	—
	0.03	0.03	0.02			

Degrees of Freedom	4	5	13,922	7	13,924	14,320
F-Stat p-value	3.3%	1.6%	0.0%	50.4%	0.0%	0.0%
Adj. R^2	71.7%	75.0%	14.4%	6.4%	3.8%	3.3%

Green: coefficient sign agrees with project level ;
Red: coefficient sign disagrees with project level
Confidence Level of Coefficient Significance: *90%; **95%; ***99%

Table 1: Summary Statistics

To provide context for the discussion of results at the subproject and task levels, as well as further tests, a short interpretation of the Full Model coefficients at the project level will be provided:

1. **Logic:** measures the proportion of tasks are missing logic. Any value over 5% is considered a flag. One would expect a positive coefficient, because the more missing logic the poorer the schedule network and less realistic the forecast. The estimated coefficient was positive and agreed with expectations. It is not significant at the 90% confidence level.
2. **Leads:** measures the proportion of tasks that have leads, which forces a task to start before its logical predecessor. Any value greater than zero is a flag. One would expect a positive coefficient for leads as a higher proportion of leads create execution risk. The estimated coefficient was positive and agreed with expectations. It is not significant at the 90% confidence level.
3. **Lags:** measures the proportion of tasks that have lags, which forces a successor task to wait before it can start. Any value greater than 5% is a flag. One would expect a positive coefficient for lags as a higher proportion of lags means that the schedule is not adequately detailed. The estimated coefficient was positive and agreed with expectations. It is significant at the 90% confidence level.
4. **Relationship Types:** measures the proportion of logic links that are Finish-to-Start. Any value less than 95% is a flag. One would expect a negative coefficient for this test as a higher proportion of Finish-to-Start relations means a less risky schedule as it prohibits concurrency. The estimated coefficient is positive and **disagreed with expectations**. It is significant at the 90% level.
5. **Hard Constraints:** measures the proportion of tasks that have hard constraints, which doesn’t allow the effects of schedule change to propagate through the schedule. Any value above 5% is a flag. One would expect a positive coefficient, as the more hard constraints there are, the less realistic the forecast will be because information is not propagated through the schedule. The estimated coefficient is positive and agrees with expectations. It is not significant at the 90% level.
6. **High Float:** measures the proportion of tasks that have high float, which may signal missing logic or too much margin built into the schedule. Any value above 5% is a flag. Though one might expect a positive coefficient, because higher values are considered to reflect poor quality, a negative coefficient is more likely. Because the first check accounts for logic, the coefficient is more likely picking up

on excess margin built into the schedule which, all else equal, would lead to a lower MAPE. The estimated coefficient on High Float is negative and agrees with expectations. It is not significant at the 90% confidence level.

7. **Negative Float:** measures the proportion of tasks that have negative float, which signals a negative effect on schedule and require management mitigation. Any value above zero is a flag. One would expect a positive coefficient because numerous tasks with negative float means the schedule will slip unless the risks are mitigated. The estimated coefficient on High Float is positive and agrees with expectations. It is significant at the 95% confidence level.
8. **High Duration:** measures the proportion of tasks that have high duration, which signals a lack of sufficient detail. Any value above 5% is a flag. One would expect a positive coefficient because numerous tasks with high duration means the schedule was not planned to an adequate level. The estimated coefficient is positive and agrees with expectations. It is significant at the 95% confidence level.
9. **Invalid Dates:** measures the proportion of tasks that have invalid or illogical date (such as an actual finish in the future). Any value above zero is a flag. One would expect a positive coefficient because numerous invalid dates signal a problem with basic schedule competency. The estimated coefficient is positive and agrees with expectations. It is significant at the 90% confidence level.
10. **Resources:** measures the proportion of tasks that have resources, which signals superior planning. Though a threshold value is not provided, lower values are flags. One would expect a negative coefficient because numerous tasks with resources means the schedule had strong planning. The estimated coefficient is positive and **disagrees with expectations**. It is not significant at the 90% confidence level.
11. **Missed Activities:** measures the proportion of tasks that were planning to be finished and have experienced slip. Any value above 5% is a flag. One would expect a positive coefficient because a large number of tasks who have missed their baseline date means the project is executing poorly. The estimated coefficient is positive and agrees with expectations. It is not significant at the 90% confidence level.

12. **Critical Path Test:** this value equals 1 if the IMS passed the Critical Path Test and 0 if it did not. Not passing the Critical Path Test is a flag. One would expect a negative coefficient because an IMS which passes the test will allow updated information to properly propagate through the entire schedule. The estimated coefficient was negative, though not significant at the 90% confidence level.
13. **Critical Path Length Index (CPLI):** measure of schedule efficiency required to finish on time. Any value below 0.95 is a flag. One would expect a positive coefficient as projects requiring less efficiency to finish on time should also be relatively more accurate and timely. The estimated coefficient is positive and agrees with expectations. It is significant at the 90% confidence level.
14. **Baseline Execution Index (BEI):** measures the proportion of tasks that have actually been finished to those which were baselined to be finished. Any value below 0.95 is a flag. One would expect a negative coefficient because executing more tasks than baselined means the project is likely ahead of schedule. The estimated coefficient is negative and agrees with expectations. It is not significant at the 90% confidence level.

- **Intercept:** the intercept represents the predicted MAPE where all values for the 14-Point Assessment equal to zero. In some cases, this says contradictory things about the schedule: a zero value for BEI means terrible performance while a zero value for Logic means good networking. Further, at the project level the CPLI does not vary far from “1”, particularly for initial schedule submissions, making its coefficient interact strongly with the intercept.

Estimated coefficients for two of the 14 checks had the opposite sign of what was expected: Relationship Types and Resources. The latter may be expected because though resource-loading is a best practice, it does not interact with the networking logic or gauge performance and thus it does not directly affect the realism of the IMS duration forecast. The former cannot be reconciled. It is expected that Finish-to-Start relationships drive schedule realism as it forces a task to finish before its successors start, implying less concurrency of tasks.

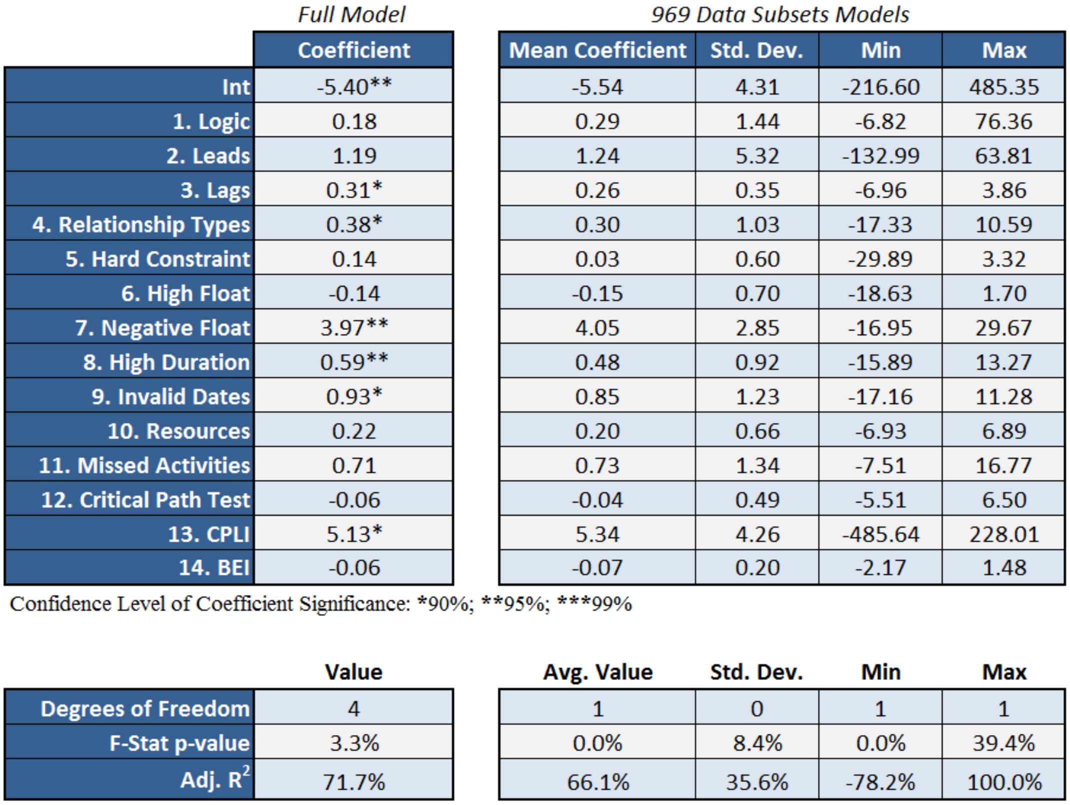
In Table 1 above, the subproject and task level coefficients were colored either green if the sign agreed with the project level coefficient in that specification, or red if they did not agree. Regressions were also performed using fixed effects for contract, as subprojects and tasks with a contract may all behave similarly. Because none of the coefficients on the variables flipped signs, the fixed effects models were dropped.

In the 13-Point Assessment specification, 5 of 14 subproject coefficients (including the intercept) disagreed on the sign relative to the project level. Two of the disagreements related to the two coefficients at the project level which did not conform to expectations: Relationship Types and Resources. The remaining three deviations were: 2. Leads; 3. Lags; and 14. BEI. Though no relevant explanation can be made for a higher proportion of Leads being correlated with a lower MAPE, there may be a plausible one for Lags. Lags signal a lack of detail as a successor activity such usually take that void. Whereas at the project level, tasks not reflected in the IMS will inevitably affect its forecast, an undetailed task in a subproject may have negative effects in other parts of the IMS, but not the subproject in question.

Of the five deviations, the BEI is the most concerning because it is significant above the 99% confidence level and it doesn't conform to expectations at the project level. It would appear, however, that the BEI in the subproject indicates not so much performance, as intended at the project level, but distance from the first IMS "time now." Subprojects which were entirely forecasts as of the first IMS all received a BEI of 1.00, as one cannot fall behind on tasks which have not started. Subprojects which are already in execution (31% of the total) had an average BEI of 0.605, far less than the 1.00 score that all future subprojects received. The coefficient for the BEI is positive at the subproject level because those subprojects in the future will realize the slip not only due to its internal execution, but the execution of all its predecessors. Predecessor uncertainty has already been realized for the most part in near-term tasks. Therefore, future subprojects will tend to have higher MAPEs because on average they have more favorable BEI scores.

In the 11-Point Assessment specification, 7 of the 11 subproject coefficients and 4 of the 12 task level coefficients disagreed with the project level. A point-by-point variance analysis will not be provided here, but it will be noted that all three models in the 11-Point Assessment specification only poorly explained variation in the MAPE, with adjusted R² values at below 7% for the project level and 4% for the subproject and task levels. Because of the far greater number of observations in the latter two levels, the F-Stat p-value, which measures the joint probability that all variable coefficients are zero, is near 0.0%, while for the project level it is at 50%. Removing the CPLI and BEI from the project level, then, has a substantial effect on the specification's overall significance.

In order to further test the stability of the coefficients, the models were tested against varying subsets of the data (cross-validation). At the project level, coefficients were estimated for all 969 possible combinations of 16 project level observations (leaving one degree of freedom). Table 2 below displays the summary information from the project level subset analysis. The means across all subset regression coefficients, in general, were very close to the estimated coefficients from the Full Model. However, the mean hides extreme variation across individual subset regression runs. Each and every estimated variable had a positive and negative coefficient value in one subset model run or another (see Min and Max columns). For all but three variables, the standard deviation of the coefficient estimates was larger than the mean.



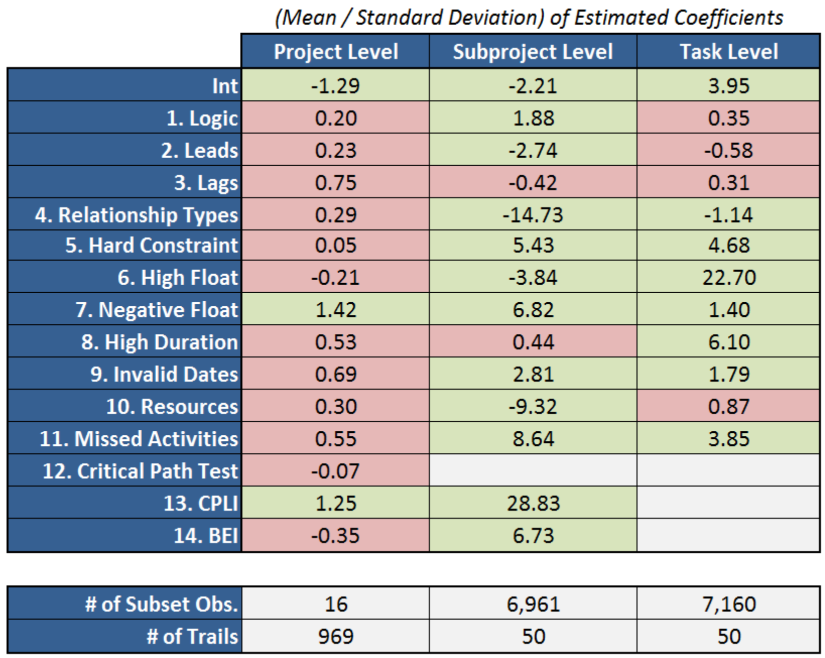
	Full Model	*969 Data Subsets Models*			
	Coefficient	Mean Coefficient	Std. Dev.	Min	Max
Int	-5.40**	-5.54	4.31	-216.60	485.35
1. Logic	0.18	0.29	1.44	-6.82	76.36
2. Leads	1.19	1.24	5.32	-132.99	63.81
3. Lags	0.31*	0.26	0.35	-6.96	3.86
4. Relationship Types	0.38*	0.30	1.03	-17.33	10.59
5. Hard Constraint	0.14	0.03	0.60	-29.89	3.32
6. High Float	-0.14	-0.15	0.70	-18.63	1.70
7. Negative Float	3.97**	4.05	2.85	-16.95	29.67
8. High Duration	0.59**	0.48	0.92	-15.89	13.27
9. Invalid Dates	0.93*	0.85	1.23	-17.16	11.28
10. Resources	0.22	0.20	0.66	-6.93	6.89
11. Missed Activities	0.71	0.73	1.34	-7.51	16.77
12. Critical Path Test	-0.06	-0.04	0.49	-5.51	6.50
13. CPLI	5.13*	5.34	4.26	-485.64	228.01
14. BEI	-0.06	-0.07	0.20	-2.17	1.48

Confidence Level of Coefficient Significance: *90%; **95%; ***99%

	Value	Avg. Value	Std. Dev.	Min	Max
Degrees of Freedom	4	1	0	1	1
F-Stat p-value	3.3%	0.0%	8.4%	0.0%	39.4%
Adj. R^2	71.7%	66.1%	35.6%	-78.2%	100.0%

Table 2: Project Level Subset Summary Information

At the subproject and task levels, 50 regressions were performed using a random selection from half the total observations. There was far less variation in the estimated coefficients at these two levels relative to the total project level. At the subproject level, all coefficients but two had standard deviations smaller than their mean values, and at the task level all but four. In most cases, the standard deviations were many times smaller than the mean, providing relative confidence in the coefficient values. See Table 3 below.



(Mean / Standard Deviation) of Estimated Coefficients

	Project Level	Subproject Level	Task Level
Int	-1.29	-2.21	3.95
1. Logic	0.20	1.88	0.35
2. Leads	0.23	-2.74	-0.58
3. Lags	0.75	-0.42	0.31
4. Relationship Types	0.29	-14.73	-1.14
5. Hard Constraint	0.05	5.43	4.68
6. High Float	-0.21	-3.84	22.70
7. Negative Float	1.42	6.82	1.40
8. High Duration	0.53	0.44	6.10
9. Invalid Dates	0.69	2.81	1.79
10. Resources	0.30	-9.32	0.87
11. Missed Activities	0.55	8.64	3.85
12. Critical Path Test	-0.07		
13. CPLI	1.25	28.83	
14. BEI	-0.35	6.73	

# of Subset Obs.	16	6,961	7,160
# of Trails	969	50	50

Green: One standard deviation away from the mean does not include zero
Red: One standard deviation away from the mean includes zero

Table 3: Coefficient Variability across Subsets

The stability of the estimated subproject and task level coefficients is expected because they have more observations, and consequently much higher degrees of freedom. The analysis suggests the estimated coefficients at the project level, even for the Full Model using all 19 observations, are not reliable. One could very well imagine that the set of 19 contracts collected for this study as one particular subset of an even larger dataset. Though estimated coefficients should be more stable across regressions of 19 observations than 16, there will still be substantial outlier subsets relative to the estimated coefficients from a large dataset (say, greater than 40).

The only coefficients in the full model which were relatively stable were Negative Float and the CPLI. Of the total variation in the MAPE explained by the Full Model at the project level (Unadjusted R^2 = 93%), a model including only the Negative Float and the CPLI explained half of that variation (Unadjusted R^2 = 46%). Despite 12 less variables to help minimize the variation in the MAPE, the R^2 unadjusted for degrees of freedom was relatively high. Negative Float and CPLI together hold about half of the total signal that the 14-Point Assessment provides concerning the IMS MAPE.

Reflections

At first glance, the 14-Point Assessment of the first IMS submission appears a relatively good predictor of the mean absolute percent error. Again, the adjusted R^2, or percent of total variance explained in the Full Model, is 71.7%, and the F-Stat p-value, or joint probability that all coefficients are zero, is 3.3%. 12 out of 14 estimated coefficients have the sign that can be expected. Those statistics, however, would seem implausible. Cross-validation using data subsets show the project level coefficients to be highly unstable. Subproject and task level model coefficients did not reflect those from the project level, but were internally consistent across data subsets. The lower-level models suggest that a large sample of project level IMSs can return consistent coefficients, but the total amount of variation explained in schedule forecast accuracy and timeliness may be relatively low.

The MAPE is a measure that takes into account both the accuracy and timeliness of schedule forecasts. So long as realized schedule quality is measured using metrics like the MAPE, one should expect the 14-Point Assessment to only explain a small portion of schedule quality variation for two reasons. First, the accuracy of an initial schedule estimate depends largely on factors that are not functions of schedule quality per se. There is a large literature on the causes of schedule slip to initial estimate, such as requirements stability, assumption realism, and integration risk.[4] Second, the timeliness of the IMS does not only depend on how quickly the IMS incorporates project

[4] For an overview, see Jessie Riposo, Megan McKernan, Chelsea Kaihoi Duan. 2014. *Prolonged Cycle Times and Schedule Growth in Defense Acquisition: A Literature Review*.

information, but when that information becomes available. For example, one contract schedule slip may be driven by requirements instability, which tends to manifest itself as a problem early, and another contract schedule slip may be driven by integration risk, which will not be apparent until relatively late.

This paper has gauged schedule quality using the MAPE, but the measure is clearly not adequate due to confounding factors. The problem of targeting schedule quality itself as distinct from all other project information does not mean the analyst can say nothing. A better definition of schedule quality is the effective incorporation of localized project knowledge into an activity-based network. The definition focuses attention on the CAMs, or those with project knowledge, instead of the scheduler who synthesizes that knowledge. A schedule's quality can only be as good as the duration estimates and activity networking, information that can only emanate from CAMs. The incentives and constraints facing CAMs that may affect the viability of schedules will not be explored. However, a short framework for understanding schedule quality will be offered and recommendations provided. First, for which project types are networked schedules not suitable? Second, for projects suitable to networked scheduling, how well does the 14 Point Assessment measure quality?

Scheduling is predominated by the networked, or algorithmic, scheduling approach which may not be preferred for all project types. "The algorithmic approach usually requires a mathematical formulation of the problem which includes objectives and constraints"[5] used to compute probabilistic project end dates. Networked schedules provide advantages only when "the problem allows for a crisp and precise mathematical formulation [and] the amount of randomness in the environment is minimal."[6] Not many CAMs in defense acquisitions would associate their work-scope with such predictability. Reputable scheduling guidance, though making brief notes of the disadvantages, have not identified which project types may require simplified or alternative approaches. Guidance fails even further to specify mitigation techniques for project types where strong logic-based networked schedules are less desirable.

[5] Pinedo and Yen, pp. 2

[6] Ibid.

Criticisms of the algorithmic approach, and quality checks pertaining to them, may apply to early research and design (R&D) projects.

Unlike mature production processes which tend to require numerous repetitive tasks that interact deterministically, R&D tends to be an open-ended system where interactions are non-deterministic and cannot be rendered plain in all cases. For example, a particular task may not have clear exit criteria because uncertain future developments in successor tasks may impact the evaluation of the former. Many tasks then progress without pre-specifiable feedback, creating reflexive instead of causal relationships. Because a design task may be dependent upon a host of other tasks, which themselves depend on the outcome of the first task, a networked schedule fails to capture the stochastic changes and instead falsely represents task definition and relationships as "crisp and precise." Yet CAMs find themselves required to provide project information in a crisp and precise manner even when circumstances do not call for it. In fact, the false representation could prove more misleading than pure reckoning by knowledgeable insiders.

It has been long recognized that the distinctive needs of R&D gets subsumed in procurement organizations devoted to production operations and standards. Just after World War II, Vannevar Bush, Director of the Office of Scientific Research and Development, addressed the president. "Research is the exploration of the unknown and is necessarily speculative. It is inhibited by conventional approaches, traditions, and standards. It cannot be satisfactorily conducted in an atmosphere where it is gauged and tested by operating or production standards."[7] The debate over separating R&D from production continues to be debated by Congressional leaders when discussing defense acquisition reform.[8] Yet the EVMS production standard, which originated in the 1960s during the McNamara years as Secretary of Defense, is not put on R&D budget activities 6.1 and 6.2 for basic and applied research. Though EVMS is often required for large R&D activities for budget activity 6.3, advanced technology development, and more mature phases, many defense programs by this stage reflect engineering and not scientific challenges.

[7] Bush.
[8] Clark.

The argument that large development programs are primarily engineering-based has deep roots. Wernher von Braun, Chief of Army missiles in 1958, said "I believe an established missile program, like the Jupiter, has much more similarity with an industrial planning job than with a scientific project […] I would say it was 90% engineering and 10% scientific."[9] Many today might have similar attitudes towards development programs. Is the networked schedule applicable to these engineering cases?

In their 1962 classic, "The Weapons Acquisition Process: An Economic Analysis," Peck and Scherer had more to say about the Jupiter Intermediate Range Ballistic Missile (IRBM) experience. They show that the engineering activities were characterized by trial-and-error processes.

> "There remained, as General Schriever noted, one critical problem – re-entry of the warhead into the atmosphere – about which little physical knowledge existed […] Even then, however, it turned out that the re-entry problem was resolved by [engineering] activities before a complete [scientific] understanding existed. The Jupiter IRBM nose cone problem was solved largely in an empirical manner. It was known from theoretical calculations that the nose cone had to resist certain general heats and shock waves. Guided by test data on rocket throat temperatures, *one material after another and one shape after another* were tried in the exhaust blast of a rocket engine until the most successful combination was found.
>
> This nose cone illustration reflects a broader set of technical problems typifying advanced weapons developments. Fundamental scientific knowledge about the environments within which new aircraft, guided missiles, and space vehicles must operate has frequently been lacking during many developments of the 1950-1960 era. For example, science has yet to provide sufficient understanding of how objects behave in various supersonic and hypersonic environments to predict fully the problems which will be encountered in flight. All too often, these problems do not become apparent until a prototype vehicle is test-flown unsuccessfully. Then isolating the problem requires *lengthy trial-and-error testing in which scientific theory may be of little assistance*."[10] [Emphasis the author's]

Many DoD programs solve technical problems with engineering trial-and-error and not by establishing scientific knowledge first to guide stable plans. Networked

[9] Peck and Scherer, pp. 40.
[10] Peck and Scherer, pp. 40.

schedules, and EVMS, provide little predictive ability in projects that require trial-and-error. The project can get stuck in activity loops without indication on when successful emergence will occur. The specification that actually works, or will provide other information on what might work, is a highly random event from the planner's perspective. The networked schedule requires stable plans and well-defined interactions. Today, R&D projects, especially large scale projects, avoid the trial-and-error approach and attempt to rely on fine planning that requires accurate forecasts of activity cost and durations. R&D projects continue to be judged according to production standards.

Figure 4 below depicts the suitability of networked scheduling through the acquisition life-cycle. The 14-Point Assessment treats all schedules as though they were production-like regardless of the disposition of project work-scope. One should account for considerations of uncertainty whenever assessing schedule quality. For example, high task durations for R&D projects may be preferred (so long as their finish dates aren't correlated), or a high proportion of relations may not be Finish-to-Start to allow for concurrence and reflexivity.

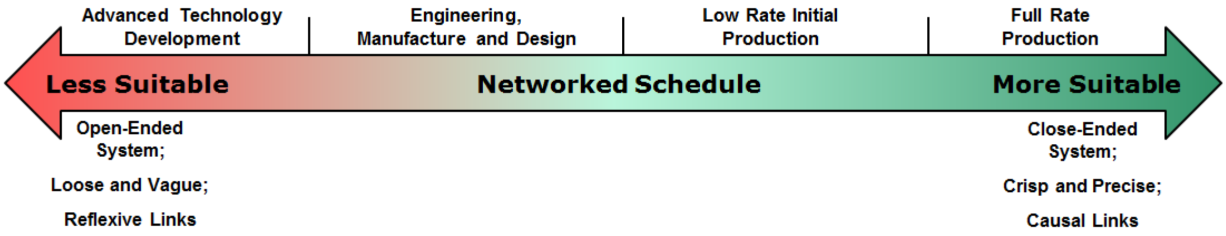


Figure 4: Algorithmic Approach in the Acquisition Life-Cycle

Even among more production-like projects where the networked schedule is suitable, the 14-Point Assessment misses important margins of quality. Regardless of how well the initial IMS synthesizes realistic CAM knowledge, once the project starts knowledge grows and tends toward disorder, similar to entropy. To understand the impact, one must recognize that a schedule's complexity is not measured so much by the number of activities, but the number of relationships. Activities should have at least one successor, often more. Where a 100 activity schedule may have 500 relationship pairs, a

1,000 activity schedule may have 5,000 relationship pairs. Note that schedule complexity as measured by number of relationships grows nonlinearly with respect to number of activities. In reality, the actual number of relationships affecting each activity increases the more detailed a schedule defines them. The concept is similar to the "square-cube" law that states that an object's volume increases faster than its surface area.

Analysts generally assume schedule complexity is invariant to scale and therefore not a primary determinant of schedule quality. The 14-Point Assessment often takes ratios of important measures to normalize for project scale. In fact, it often favors detail that increases schedule size. Yet the practice hides those schedule quality signals affected by scale. Suppose a schedule could be equally represented by 100 high-duration activities with 500 relations and 1,000 activities with 5,000 relations. The former schedule would be flagged for high-durations, and quality assessed to be lower than the latter. Such a conclusion could be erroneous for a couple reasons.

First, the likelihood that real activity relationships remain undefined increases with scale. For example, consider an individual activity that can be further separated into 10 activities. Not only will each lower-level activity have to correctly incorporate the relations between each other, but from those of its parent to outside activities. The margin for error in defining relations increases. Further, scope of each lower-level activity is necessarily more uncertain than that for the parent, increasing the likelihood of either a change to the activity or its removal and replacement. A detailed schedule can actually abstract away from real project information more than a high-level schedule. Where relationships go undefined, new project information will not reverberate through the schedule realistically. The result often suppresses forecast volatility in the short-run but creates big jumps, or surprises, later on.

Second, developing networked schedules is "difficult and time consuming [and] only as sound as the activity time and resource estimates."[11] In practice, increasing schedule scale requires not only vastly more effort by the CAMs to plan, but to track actuals and incorporate new information. Determining the effectiveness of schedule

[11] Defense Acquisition University (DAU), "DAU Scheduling Guide," pp. 41.

maintenance cannot be performed by viewing the current submission in a vacuum. Longitudinal, or cross-submission, quality checks provide key signals that the 14-Point Assessment overlooks. For example, the 100 activity schedule with high-duration tasks gets flagged, but its ability to maintain coherence across time is far superior to the 1,000 activity schedule. Where a static 14-Point Assessment alone is implemented, the 1,000 activity schedule could continually pass by reconfiguring parameters and baselines for each submission. Some relevant longitudinal checks that can detect such reconfigurations include the proportion of tasks entering or exiting the schedule, the extent of baseline changes, execution to short-term plan, proportion of relationships changed, etc.

Like the 14-Point Assessment, the longitudinal checks need not be complicated to be effective. In fact, such heuristics side-step the problem of measuring schedule quality because their common-sense logic does not require external validation. Passing scores on heuristics merely point to an absence of evidence of poor schedule quality, and does not provide positive evidence of good schedule quality. The concept is similar to a trial by jury, where the defendant can only be found “guilty” or “not guilty,” but never “innocent.” It can be shown that the defendant did the crime, but lacking evidence of guilt does not prove innocence. It creates a reasonable doubt about the guilty charge. With regards to schedule quality, it can be proved that the CAM ineffectively incorporated local project knowledge, such as by not providing inter-task relationships, but it cannot be proved that he did so effectively. The more margins the analyst can check, the more confident the analyst can be that an absence of evidence of poor quality is good enough to declare “fair” or even “high” quality.

The key point is that important evidence of quality is only available through an analysis of the schedule as a dynamic process. If one believes that longitudinal checks are important to assessing margins of schedule quality not picked up by the static 14-Point Assessment, then their inclusion can go a long way to increasing confidence. Say an analyst, using the 14-Point Assessment, is 70% confident that a schedule is good enough for stakeholders to base decisions upon. If another analyst runs the same static checks and is also 70% confident of the adequacy of its quality, then the overall confidence is still 70% because they are analyzing the same set of information. But, if the other analyst runs longitudinal checks and says that he is 70% confident on those quality margins, then the overall confidence of schedule quality is *greater than 70%.*[12] The phenomenon occurs where distinctive sets of information pertaining to the same question are brought to bear. The point is not to seek more of the same kinds of data, but more of different kinds of data. If longitudinal checks are important measures of quality, adding them to an overall assessment will increase the confidence one has in determining the true quality.

There may be further margins on which to gauge schedule quality, and they too should be added if discovered. For example, maintaining a parallel schedule will allow the analyst to observe *ex post* relative, though not absolute, schedule quality. Yet even in project execution managers may intuit which version is superior. Other parallel duration forecasts can come from knowledge-based approaches[13] and internal prediction markets.[14] Where possible to run parallel duration forecast techniques, valuable additional schedule quality information can be generated. Again, the analyst doesn't necessarily need more checks, but different checks. The fact that multicollinearity was observed in the parametric analysis suggests that some of the 14-Points might be picking up the same variation.

In practice, the IMS will often be the only set of cohesive project information related to duration. It is preferred that the IMS exhibits high quality such that the analyst can develop actionable plans using its forecast. Currently, analysts spend a great deal of effort trying to tease out the biases in schedules to generate more "realistic" forecasts. Such manipulation of schedule data is highly speculative. Attention to schedule maintenance is a superior use of scarce resources relative to guessing its inherent biases. Further, the longitudinal checks would actually expose the need to eliminate the biases that analysts attempt to exploit for the purpose of prediction.[15]

[12] For a discussion of this concept, see Tetlock and Gardner, *Superforecasting: The Art and Science of Prediction.*

[13] Duchessi and O'Keefe, "A Knowledge-Based Approach to Production Planning." *The Journal of the Operational Research Society*, Vol. 41, No. 5, Artificial Intelligence and Expert Systems. Part I. Production Planning, Applications and Methodology (May, 1990), pp. 377-390

[14] Hanson, 2008.

[15] The forecasting technique done in the author's "Trust, but Verify" paper is a basic application of longitudinal quality checks. It uses the maximum observed activity slip relative to the initial IMS parameters to calculate the

Conclusion

This paper has cast suspicion upon the ability of the 14-Point Assessment to predict a schedule forecast's accuracy and timeliness. Ultimately, the quality of a schedule cannot be measured as distinct from project assumptions and performance. However, by using common-sense heuristics, one can expect a fair gauge by systematically searching for evidence of poor quality. Yet the 14-Point Assessment misses several important margins of schedule quality, namely the ability of the schedule to evolve consistently and incorporate new information reliably. Simple longitudinal checks are advocated to increase the confidence one has in a schedule quality assessment. Other independent data should be considered where possible. The expansion of basic heuristics should provide large returns by eliminating obvious methods for a poor schedule to pass static 14-Point Assessments. Further study is required on exactly which longitudinal checks provide the best value. Additionally, studies on where the heuristics should be flexible with respect to project type (e.g., R&D) and alternative schedule approaches are advocated.

impact on total schedule. Making schedulers aware of such biases would go a long way to eliminating the technique's ability to forecast.

Bibliography:

Bridgeforth, Shedrick. “Using Earned Value Data to Forecast the Duration of Department of Defense (DoD) Space Acquisition Programs.” 26 Mar 2015. http://www.dtic.mil/docs/citations/ADA615411

Bush, Vannevar. 1945. “Science – The Endless Frontier.” http://www.nsf.gov/about/history/vbush1945.htm

Clark, Colin. 2016. “Thornberry’s Buying Bill Adds Bureaucracy, Helps Biz With IP.” http://breakingdefense.com/2016/03/thornberrys-buying-bill-adds-bureaucracy-helps-biz-with-ip/

DCMA, Defense Contract Management Agency. “Earned Value Management System (EVMS) Program Analysis Pamphlet (PAP). DCMA-EA PAM 200.1.” 2012.

Defense Acquisition University (DAU), “DAU Scheduling Guide,” https://acc.dau.mil/CommunityBrowser.aspx?id=37441&lang=en-US.

Duchessi, P. and O’Keefe, R.M. “A Knowledge-Based Approach to Production Planning.” *The Journal of the Operational Research Society*, Vol. 41, No. 5, Artificial Intelligence and Expert Systems. Part I. Production Planning, Applications and Methodology (May, 1990), pp. 377-390.

Hanson, Robin. “The Promise of Prediction Markets, Science, Vol. 320, No. 5878 (May, 2008).

Jessie Riposo, Megan McKernan, Chelsea Kaihoi Duan. 2014. *Prolonged Cycle Times and Schedule Growth in Defense Acquisition: A Literature Review*. RAND. http://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR455/RAND_RR455.pdf

Peck, Martin and Scherer, Fredric. 1962. “The Weapons Acquisition Process: An Economic Analysis.” Harvard Business School, 1962.

Pinedo, Michael and Yen, Benjamin. “On the Design and Development of Scheduling Systems”, 2014, https://www.researchgate.net/publication/2284490_On_the_Design_and_Development_of_Scheduling_Systems.

Tetlock, Philip and Gardner, Dan. “Superforecasting: The Art and Science of Prediction.” Crown, 2015.

Appendix: 14-Point Assessment

1. Logic

This metric identifies incomplete tasks with missing logic links. It helps identify how well or poorly the schedule is linked together. Even if links exist, the logic still needs to be verified by the technical leads to ensure that the links make sense. Any incomplete task that is missing a predecessor and/or a successor is included in this metric. The number of tasks without predecessors and/or successors should not exceed 5%. An excess of 5% should be considered a flag.

Project and Subproject Level:

$$Missing\ Logic\ \% = \frac{\#\ of\ tasks\ missing\ logic}{\#\ of\ incomplete\ tasks}\ x\ 100$$

Task Level:

$$Missing\ Logic = 1$$

$$Not\ Missing\ Logic = 0$$

2. Leads

This metric identifies the number of logic links with a lead (negative lag) in predecessor relationships for incomplete tasks. The critical path and any subsequent analysis can be adversely affected by using leads. The use of leads distorts the total float in the schedule and may cause resource conflicts. Per the IMS Data Item Description (DID), negative time is not demonstrable and should not be encouraged. Using MS Excel, count the number of “Leads” that are found. Leads should not be used; therefore, the goal for this metric is 0.

Project and Subproject Level:

$$Leads\ \% = \frac{\#\ of\ logic\ links\ with\ leads}{\#\ of\ logic\ links}\ x\ 100$$

Task Level:

$$One\ or\ more\ leads = 1$$

$$No\ leads = 0$$

3. Lags

This represents the number of lags in predecessor logic relationships for incomplete tasks. The critical path and any subsequent analysis can be adversely affected by using lags. Per the IMS DID, lag should not be used to manipulate float/slack or to restrain the schedule. Using MS Excel, count the number of "Lags" that are found. The number relationships with lags should not exceed 5%.

Project and Subproject Level:

$$Leads\ \% = \frac{\#\ of\ logic\ links\ with\ lags}{\#\ of\ logic\ links}\ x\ 100$$

Task Level:

$$One\ or\ more\ lags = 1$$

$$No\ lags = 0$$

4. Relationship Types

The metric provides a count of incomplete tasks containing each type of logic link. The Finishto-Start (FS) relationship type ("once the predecessor is finished, the successor can start") provides a logical path through the program and should account for at least 90% of the relationship types being used. The Start-to-Finish (SF) relationship type is counter-intuitive ("the successor can't finish until the predecessor starts") and should only be used very rarely and with detailed justification. By counting the number of Start-to-Start (SS), Finish-to-Finish (FF), and Start-to-Finish (SF) relationship types, the % of Finish-to-Start (FS) relationship types can be calculated.

Project, Subproject, and Task Level:

$$\%\ of\ FS\ Relationship\ Types = \frac{\#\ of\ logic\ links\ with\ FS\ Relationships}{\#\ of\ logic\ links}\ x\ 100$$

5. Hard Constraints

This is a count of incomplete tasks with hard constraints in use. Using hard constraints [Must-Finish-On (MFO), Must-Start-On (MSO), Start-No-Later-Than (SNLT), & Finish-No-Later-Than (FNLT)] may prevent tasks from being moved by their dependencies and, therefore, prevent the schedule from being logic-driven. Soft constraints such as As-Soon-As-Possible (ASAP), Start-No-Earlier-Than (SNET), and Finish-No-Earlier-Than (FNET) enable the schedule to be logic-driven. Divide the total number of hard constraints by the number of incomplete tasks. The number of tasks with hard constraints should not exceed 5%.

Project and Subproject Level:

$$Hard\ Constraint\ \% = \frac{Total\ \#\ of\ incomplete\ tasks\ with\ hard\ constraints}{Total\ \#\ of\ incomplete\ tasks}\ x\ 100$$

Task Level:

$$One\ or\ more\ hard\ constraints = 1$$

$$No\ hard\ constraints = 0$$

6. High Float

An incomplete task with total float greater than 44 working days (2 months) is counted in this metric. A task with total float over 44 working days may be a result of missing predecessors and/or successors. If the percentage of tasks with excessive total float exceeds 5%, the network may be unstable and may not be logic-driven.

Project and Subproject Level:

$$High\ Float\ \% = \frac{Total\ \#\ of\ incomplete\ tasks\ with\ high\ float}{Total\ \#\ of\ incomplete\ tasks}\ x\ 100$$

Task Level:

$$(Float \geq 44\ days) = 1$$

$$(Float < 44\ days) = 0$$

7. Negative Float

An incomplete task with total float less than 0 working days is included in this metric. It helps identify tasks that are delaying completion of one or more milestones. Tasks with negative float should have an explanation and a corrective action plan to mitigate the negative float. Divide the total number of tasks with negative float by the number of incomplete tasks. Ideally, there should not be any negative float in the schedule.

Project and Subproject Level:

$$Negative\ Float\ \% = \frac{Total\ \#\ of\ incomplete\ tasks\ with\ negative\ float}{Total\ \#\ of\ incomplete\ tasks}\ x\ 100$$

Task Level:

$$(Float < 0\ days) = 1$$

$$(Float > 0\ days) = 0$$

8. High Duration

An incomplete task with a baseline duration greater than 44 working days (2 months), and has a baseline start date within the detail planning period or rolling wave is included in this metric. It helps to determine whether or not a task can be broken into two or more discrete tasks rather than one. In addition, it helps to make tasks more manageable; which provides better insight into cost and schedule performance. Divide the number of incomplete tasks with high duration tasks by the total number of incomplete tasks. The number of tasks with high duration should not exceed 5%.

Project and Subproject Level:

$$High\ Duration\ \% = \frac{Total\ \#\ of\ incomplete\ tasks\ with\ high\ duration}{Total\ \#\ of\ incomplete\ tasks}\ x\ 100$$

Task Level:

$$(Duration \geq 44\ days) = 1$$

$$(Duration < 44\ days) = 0$$

9. Invalid Dates

Incomplete tasks that have a forecast start/finish date prior to the IMS status date, or has an actual start/finish date beyond the IMS status date are included in this metric. A task should have forecast start and forecast finish dates in the future relative to the status date of the IMS (i.e. if the IMS status date is 8/1/09, the forecast date should be on or after 8/1/09). A task should not have an actual start or actual finish date that is in the future relative to the status date of the IMS (i.e. if the IMS status date is 8/1/09, the actual start or finish date should be on or before 8/1/09, not after 8/1/09). There should not be any invalid dates in the schedule.

10. Resources

This metric provides verification that all tasks with durations greater than zero have dollars or hours assigned. Some contractors may not load their resources into the IMS. The IMS DID (DIMGMT-81650) does not require the contractor to load resources directly into the schedule. If the contractor does resource load their schedule, calculate the metric by dividing the number of incomplete tasks without dollars/hours assigned by the total number of incomplete tasks.

Project and Subproject Level:

$$Missing\ Resource\ \% = \frac{Total\ \#\ of\ incomplete\ tasks\ with\ missing\ resource}{Total\ \#\ of\ incomplete\ tasks}\ x\ 100$$

Task Level:

$$Missing\ Resource = 1$$

$$Not\ Missing\ Resource = 0$$

11. Missed Tasks

A task is included in this metric if it is supposed to be completed already (baseline finish date on or before the status date) and the actual finish date or forecast finish date (early finish date) is after the baseline finish date or the Finish Variance (Early Finish minus Baseline Finish) is greater than zero. This metric helps identify how well or poorly the schedule is meeting the baseline plan. To calculate this metric, divide the number of missed tasks by the baseline count which does not include the number of tasks missing baseline start or finish dates. The number of missed tasks should not exceed 5%.

Project and Subproject Level:

$$Missed\ \% = \frac{\#\ of\ tasks\ with\ actual\ or\ forecast\ finish\ date\ past\ baseline\ date}{\#\ of\ tasks\ with\ baseline\ finish\ dates\ on\ or\ before\ status\ date}\ x\ 100$$

Task Level:

$$Actual\ or\ forecast\ finish\ past\ baseline\ date = 1$$

$$Actual\ or\ forecast\ finish\ not\ past\ baseline\ date = 0$$

12. Critical Path Test

The purpose is to test the integrity of the overall network logic and, in particular, the critical path. If the project completion date (or other milestone) is not delayed in direct proportion (assuming zero float) to the amount of intentional slip that is introduced into the schedule as part of this test, then there is broken logic somewhere in the network. Broken logic is the result of missing predecessors and/or successors on tasks where they are needed. The IMS passes the Critical Path Test if the project completion date (or other task/milestone) show a negative total float number or a revised Early Finish date that is in direct proportion (assuming zero float) to the amount of intentional slip applied.

This test was not applied to the subproject or task level.

13. Critical Path Length Index (CPLI)

The Critical Path Length Index (CPLI) is a measure of the efficiency required to complete a milestone on-time. It measures critical path "realism" relative to the baselined finish date, when constrained. A CPLI of 1.00 means that the program must accomplish one day's worth of work for every day that passes. A CPLI less than 1.00 means that the program schedule is inefficient with regard to meeting the baseline date of the milestone (i.e. going to finish late). A CPLI greater than 1.00 means the program is running efficiently with regard to meeting the baseline date of the milestone (i.e. going to finish early). The CPLI is an indicator of efficiency relating to tasks on a milestone's critical path (not to other tasks within the schedule). The CPLI is a measure of the relative achievability of the critical path. A CPLI less than 0.95 should be considered a flag and requires further investigation.

The CPLI requires determining the program schedule's Critical Path Length (CPL) and the Total Float (TF). The CPL is the length in work days from time now until the next program milestone that is being measured. TF is the amount of days a project can be delayed before delaying the project completion date. TF can be negative, which reflects that the program is behind schedule. The mathematical calculation of total float is generally accepted to be the difference between the "late finish" date and the "early finish" date (late finish minus early finish equals total float).

Project and Subproject Level:

$$Critical\ Path\ Length\ Index\ (CPLI) = \frac{CPL + TF}{CPL}$$

Task Level: N/A

14. Baseline Execution Index (BEI)

The Baseline Execution Index (BEI) metric is an IMS-based metric that calculates the efficiency with which tasks have been accomplished when measured against the baseline tasks. In other words, it is a measure of task throughput. The BEI provides insight into the realism of program cost, resource, and schedule estimates. It compares the cumulative number of tasks completed to the cumulative number of tasks with a baseline finish date on or before the current reporting period. BEI does not provide insight into tasks completed early or late (before or after the baseline finish date), as long as the task was completed prior to time now. See the Hit Task Percentage metric below for further insight into on-time performance. If the contractor completes more tasks than planned, then the BEI will be higher than 1.00 reflecting a higher task throughput than planned. Tasks missing baseline finish dates are included in the denominator. A BEI less than 0.95 should be considered a flag and requires additional investigation.

Project and Subproject Level:

$$Baseline\ Execution\ Index\ (BEI) = \frac{Total\ \#\ of\ tasks\ complete}{\begin{pmatrix} Total\ \#of\ tasks\ completed\ before\ now\ + \\ Total\ \#of\ tasks\ missing\ baseline\ finish\ date \end{pmatrix}}$$

Task Level: N/A