Scheduling Metrics and the Dangers of Remaining Silent

By Beatrice Nasui, PMP and
Ron Winter, PSP, FAACE

Introduction

Schedule metrics involves the measurement of any particular parameter in a plan of work, such as the number of activities, constraints or relationships in a schedule. While the subject of schedule metrics has been adequately covered in literature, the lack of standardization in the definition of metrics and their acceptable limits remains a source of ongoing debate.

The schedule is an essential management tool and its purpose is to predict completion dates for the project's activities and assist project teams in the decision-making process. As such, the schedule must convey the right information so the dates should be reliable. One of the initial steps in the schedule review process is the metrics evaluation that provides an easy way to verify the schedule integrity. However, a schedule that passes a metrics test proves that it meets the stated limits of the parameters tested and nothing more. There are no indications if the plan is realistic, feasible, or if it reflects the stated execution strategy. What importance does the schedule metrics evaluation play in the schedule review and approval process?

Since the introduction of the US Defense Contract Management Agency (DCMA) schedule evaluation protocol in 2005 and of its subsequent revisions, the scheduling community started taking sides on the matter of its applicability. In the years that followed, numerous papers covered the subject with views ranging from praising and recommending the methodology to those highlighting its numerous flaws. Most reviews merely explained how to perform the analysis without actually weighing-in on its appropriateness or limitations.

In a 2011 paper, DCMA 14-Point Schedule Assessment, this author performed a thorough evaluation of the protocol and highlighted a series of inconsistencies and implementation issues [1, pp. 20-21]. A decade later, not much seems to have changed and this analysis persists and even thrives in the industry. Now other metrics tests have entered the schedule quality analysis field.

Acumen Fuse software makes an attempt at standardization and proposes the Schedule Quality Index^{TM} as a mean to quantify the schedule quality. More recently, SmartPM^{TM} Technologies, a schedule analytics company specializing in the construction industry, introduced its own version of a schedule quality index, based on the DCMA's protocol. But while these indices become increasingly popular, their adoption needs to be made on an informed decision.

The authors believe that the original intent of using the schedule metrics was to improve the quality of the submitted schedules by adherence to predefined thresholds. The difficulty lies in defining the term “schedule quality”. This paper will highlight that there is no one common definition for all schedules so there is no one common metrics-based evaluation protocol.

A brief overview of the most popular schedule standards and guidelines is presented, followed by the introduction of software solutions for schedule metrics evaluation[1]. The subsequent section highlights the importance of using the correct metrics and corresponding thresholds for a given schedule and pinpoints the differences that exist in some of the most popular schedule checks. The authors present some of the factors that need to be considered while selecting the metrics and establishing their thresholds.

The paper addresses the importance of using the correct evaluation protocol for each situation. In addition, one needs to have a good understanding of the system’s configuration when performing the analysis using a commercial software. The paper also includes a series of recommendations for the schedule evaluation using metrics and concludes by reminding practitioners that this evaluation is only the initial step of any thorough schedule review.

Among the recent developments that gain traction in the scheduling community, the authors look into the issue of benchmarking the various schedule metrics. The benchmarking feature proposed by Deltek is becoming popular and the authors highlight some of the issues pertaining to the use of the tool.

A Literature Review of Schedule Metrics

Standards and Guidelines

Many public standards and guidelines exist that prescribe the schedule quality, with most of them having been specifically designed to evaluate detailed control schedules. These detailed control schedules, defined as Class 2 schedules in the AACE Recommended Practice 27R-03 Schedule Classification System, can exist as independent entities or be developed as parts of an Integrated Master Schedule (IMS) [2, p. 8]. The lowest level of detail that a Class 2 schedule is usually presented is a Level 4, although Level 3 schedules also are frequently used in the EPC and EPCM environment.

[1] The example used were obtained by using Deltek’s Acumen Fuse version 8.5, and Schedule Analyzer for the Enterprise, Baseline Checker version 4.

While the literature abounds in best practices recommendations and guidelines, what seems to be missing is the general consensus as to what characteristics a quality schedule should exhibit. To prove this point, the criteria included in the most popular schedule checks are listed below.

To begin, the DCMA 14-point assessment is included in the following table [3, pp. 28-32]:

No.	Criteria	Condition being evaluated	Threshold
1	Logic	Number of incomplete tasks without predecessors and/or successors	Less than 5%
2	Leads	Number of logic links with a lead (negative lag) in predecessor relationships for incomplete tasks	0
3	Lags	Number of lags in predecessor logic relationships for incomplete tasks	Less than 5%
4	Relationship types	Number of relationships for incomplete tasks	Minimum 90% Finish-to-Start (FS) relationships
5	Hard Constraints	Number of incomplete tasks with hard constraints in use	Less than 5%
6	High Float	Number of incomplete tasks with total float greater than 44 working days	Less than 5%
7	Negative Float	Number of incomplete tasks with total float less than 0 working day	0
8	High Duration	Number of incomplete tasks with a duration greater than 44 working days	Less than 5%
9	Invalid Dates	Number of incomplete tasks with forecast start/finish dates prior to the status date or with actual start/finish dates beyond the status date	0
10	Resources	Number of incomplete tasks with durations greater than zero that have dollars or hours assigned	100%
11	Missed Tasks	Number of tasks that have been completed or will finish later than planned in the baseline	Less than 5%
12	Critical Path Test	Measures the slippage of the project completion date (or other milestone) when an intentional slip is introduced in the network	Should be proportional with the intentional slip applied
13	Critical Path Length Index (CPLI)	Measures the critical path "realism" relative to the baselined finish date	Not less than 0.95
14	Baseline Execution Index (BEI)	Cumulative number of tasks completed compared to cumulative tasks with baseline finish date on or before the status date	Not less than 0.95

Table 1. DCMA-14 Point Assessment for Integrated Master Schedules (IMS)

The original DCMA-14 definition as well as numerous independent papers explain the process involved in collecting these metrics. A detailed look into the implication behind these quality checks is available in this author's paper, DCMA 14-Point Schedule Assessment [1].

A different set of quality metrics criteria is included in NASA's Scheduling Management Handbook, as shown below [4, pp. 57-59]:

No.	Criteria	Threshold			Weighting
		Green (1)	Yellow (2)	Red (3)	
1	Missing predecessors	< 5%	5-10%	> 10%	20%
2	Missing successors	< 5%	5-10%	> 10%	20%
3	Constraints and Assigned Deadlines	< 10%	10-15%	> 15%	15%
4	Tasks and Milestones Needing Status	0	0-5%	> 5%	20%
5	Actual starts/finishes after the Status Date	0	0-5%	> 5%	10%
6	Tasks marked as Milestones (but have Duration > 0)	0	0-5%	> 5%	5%
7	Summary tasks with logic ties	< 2%	2-3%	> 3%	10%
	Overall Project Rating	**> 2.5**	**1.75-2.5**	**< 1.75**	

Table 2. NASA Schedule Logic Credibility Health Check

These metrics are then scored from 1 to 3, based on the percentages found. The final score, known as Overall Project Rating, with values between 1 and 3, is determined by weighing the respective individual scores as indicated.

The US Government Accounting Office (GAO) has published a schedule quality metrics practice titled Best Practices for Project Schedules. It proposes its own 10-point checklist and associated quantitative measurements [5, pp. 183-188].

The US Naval Air Systems Command (NAVAIR) developed a similar 11-point assessment protocol as GAO, and both have a focus broader than just schedule metrics. Scope completeness, vertical and horizontal integrations of activities, resource allocations and critical path analysis complement the schedule diagnostic performed through quantitative checks [6, pp. 85-91]. But even if they include some of the more common metrics used in other protocols, their thresholds vary, as indicated in the Table 3 below:

No.	Criteria	Condition being evaluated	Target
1	Logic	Number of predecessors or successors	Minimum 1, maximum 10
2	Constraints	Soft and hard constraints	Less than 5%
3	Lags	Positive or negative	Not specified
4	High durations	Tasks with durations greater than 60 calendar days within a rolling wave (or 6 month) boundary	Less than 5%
5	"Hit or Miss" ratio analysis	Tasks completed in the past 3 months with a finish date different than baseline's	Less than 5%

Table 3. Metrics Included in NAVAIR Schedule Assessment Protocol

The importance of performing schedule health assessments is recognized by the National Defense Industrial Association (NDIA) as well, in its Planning and Scheduling Excellence Guide (PASEG). This guide lists some of the common metrics as possible choices for the analysis, namely:

- Tasks Missing Logic (predecessors and successors)
- Use of Leads (acceleration of a successor activity)
- Use of Lags (delay of a successor activity)
- Relationship Types (SS, SF, FS, FF)
- Hard Constraints (such as P6 constraints labeled, Must Finish On and Must Start On)
- High Float (High Slack)
- Negative Float (Negative Slack)
- High Task Duration
- Invalid Forecast Dates
- Resource Loading

Some new metrics are also proposed, such as:

- Over-allocated Resources, measuring the number of tasks with resources that are scheduled to occur in excess of their availability
- Tasks Without Baseline, measuring the number or percentage of tasks without baseline dates
- Out of Sequence Status, measuring the number of tasks that started earlier than the logic would allow.

However, NDIA does not define any specific thresholds, but rather recommends threshold guidelines that could trigger additional analysis [7, pp. 138-144].

Software Solutions

While all these metrics can be calculated manually in a spreadsheet, the process can prove time-consuming and is prone to errors. Many software providers responded to the increasing interest in the metrics analysis and started integrating the most referenced protocols in their products or even creating their own checks. Among the popular tools available today, Oracle's P6 schedule check functionality, available for their EPPM platform, includes its 14-point checks, as shown in the Table 4 below:

No.	Criteria	Condition being evaluated	Target
1	Logic	Activities missing predecessors or successors	Less than 5%
2	Negative Lags	Relationships with a lag duration of less than 0	Less than 1%
3	Lags	Relationships with a positive lag duration	Less than 5%

No.	Criteria	Condition being evaluated	Target
4	Long Lags	Relationships with a lag duration greater than 352h*	Less than 5%
5	Relationship Types	The majority of relationships should be Finish to Start	Greater than 90%
6	Hard Constraints	Constraints that prevent activities being moved	Less than 1%
7	Soft Constraints	Constraints that do not prevent activities being moved	Less than 5%
8	Large Float	Activities with total float greater than 352h*	Less than 1%
9	Negative Float	Activities with a total float less than 0	Less than 1%
10	Large Durations	Activities that have a remaining duration greater than 352h*	Less than 5%
11	Invalid Progress Dates	Activities with invalid progress dates	Less than 1%
12	Resource / Cost	Activities that do not have an expense or a resource assigned	Less than 1%
13	Late Activities	Activities scheduled to finish later than the project baseline	Less than 5%
14	BEI	Baseline Execution Index	Greater than 0.95

Table 4. Oracle P6 Schedule Check

*Note: the 352h label refers to 352 hours and is equivalent to 44 days (considering a working schedule of 8hrs/day).

Another software provider, Deltek through its Acumen Fuse product, makes available multiple evaluation protocols: DCMA 14-point assessment, GAO and NASA Health Check. It also includes its own schedule review criteria that produces an overall score to evaluate the schedule quality, the Schedule Index. The various metrics that compose the Schedule Index are included in the Table 5 below:

No.	Criteria	Condition being evaluated	Threshold			
			None	Low	Medium	High
1	Missing Logic	Total number of activities that are missing a predecessor, a successor, or both	Less than 5%	Less than 10%	Less than 25%	More than 25%
2	Logic Density™	The average number of logic links per activity	Minimum 2, maximum 4			
3	Critical	Number of critical activities	Point of reference	0	50%	100%
4	Hard Constraints	Number of activities with hard or two-way constraints	0	Less than 5%	5-25%	More than 25%
5	Negative Float	Number of activities with total finish float less than 0 working days	0	Less than 5%	5-25%	More than 25%
6	Insufficient Detail™	Number of activities that have a duration longer than 10% of the total duration of the project	Less than 5%	0	Less than 5%	More than 5%

No.	Criteria	Condition being evaluated	Threshold			
			None	Low	Medium	High
7	Number of Lags	Total number of activities that have lags in their predecessors	Less than 5%	Less than 25%	25-50%	More than 50%
8	Number of Leads	Total number of activities carrying negative lag	Less than 5%	Less than 25%	25-50%	More than 50%
9	Merge Hotspot	Total number of activities with a high number of predecessor links (more than 2)	0	Less than 10%	Up to 25%	More than 25%

Table 5. The Metrics that Compose the Schedule Index

Using pre-defined weightings, these metrics are then combined into an overarching Schedule Index. The value of the Schedule Index can vary from 0 to 100, with a pass/fail threshold established at the 75 mark or higher. It is suggested that any value lower than 75 should trigger a schedule rejection or at least a review and /or an update. The creator of this software, Dr. Dan Patterson suggests that if organizations are looking to outperform, they should consider raising this threshold to 85, but provides no additional information to support this recommendation [8, p. 5].

With Acumen Fuse, users can create their own evaluation protocols by selecting any of the industry standard metrics, additional metrics included in the product's libraries or their own developed metrics. This is a very powerful feature that can meet the needs of expert users, allowing them to define the metrics calculation formula, its inclusions or exclusions filters and its tripwire thresholds scales.

Another diagnostic tool, Schedule Analyzer, in its Baseline Checker Module, incorporates 2 agencies checks, DCMA and NAVFAC standards, while also performing 81 different checks including the set of checks recommended in the AACE Recommended Practice 78R-13 Original Baseline Schedule Review [9]. The Update Checker module of the same software incorporates the update portions of the DCMA and PASEG Execution Metrics checks, with 153 additional checks including the set of checks recommended in the AACE Recommended Practice 53R-06 Schedule Update Review [10]. These checks are more thorough than just looking at activity statistics. They cover reviews of calendars, calculation settings, activity steps, activity codes, expenses, resources, memos, attached documents, issues, and user-defined fields.

SmartPM, a young schedule analytics company, proposes its own Schedule Quality Index, based on the DCMA-14 point protocol. In the brief white paper that describes the index, the justification given for the selection of each of the fourteen metrics is not convincing and seems odd at times. For example, the *Total Relationships to Activities*

Ratio is presented as a metric meant to determine the structural soundness of a schedule. It then states that, to score a good grade, this ratio should be closer to 2:1, corresponding to two successors to every one activity (instead of one predecessor and one successor). Also, High Float Activities metric is presented as an indicator of flawed logic, and the authors recommend schedulers to 'do the best they can to minimize High Float Activities because schedules need to be 'reactive' to delays' [11, p. 3]. But other than the documentation that describes the metrics, users need to understand the pertinence of the proposed schedule checks, knowing the method's weaknesses. As the company caters to the construction industry, the schedule quality index would concern construction schedules, developed mainly for the execution phase.

The advances of technology are continually pushing the schedule analytics boundaries. New functionalities, new products and new players are entering the industry, trying to make best use of the schedule data. As shown in the following section, the users challenge is to select products adapted to their needs, rather than adopt tools and standards by mere convenience or commercial pressure.

Using the Appropriate Metrics and Applicable Thresholds

In the absence of contractual prescription, the problem of choosing the appropriate metrics to use from the multitude of standards and guidelines that address the schedule quality is daunting. The difficulty in reaching a consensus on schedule quality criteria is due to the fact that one single measure for any single schedule for every situation probably cannot be designated.

The schedules are developed throughout the project lifecycle and their purpose and characteristics evolve over time. This fact is reflected in the AACE Recommended Practice 27R-03, Schedule Classification System that makes the distinction between the schedule class, linked to the maturity of the information that supports the schedule development, and the schedule levels, that only reflects the schedule granularity in its presentation [2, p. 2]. Usually there is a correlation between the schedule class and the schedule level, that ensures that the schedule is developed in line with the amount and quality of the available information.

Class 5 and 4 schedules are preliminary, Level 1 or 2 schedules, developed during concept screening and pre-feasibility studies, based on limited information. The Class 3 schedule, prepared during the feasibility stage for budget authorization and / or funding, is the initial control schedule, developed usually to Level 3. The Class 2, detailed control schedule, is developed during the execution phase with a higher granularity, to Level 4.

Using any of the protocols that were designed for Level 4 baseline execution schedules on Level 3 or lower schedules would most likely generate a Fail result. This is not to say that Level 3 or lower schedules should be exempt from quality checks, only that the protocols to be used in their evaluation need to be adapted to their conditions.

In her paper "Which Schedule Quality Assessment Metrics to Use? … and When?, Shoshanna Fraizinger proposes a schedule quality metric application by schedule class [12, p. 20]. In this case, the Class 3 schedule referenced seems to reflect the baseline construction schedule approved for execution.

Other than the class that a schedule belongs to and the level to which it was developed, additional factors affect the suitability of the metrics to be retained in the analysis or their imposed thresholds. The industry for which the schedule was prepared brings its own characteristics. For example: engineering schedules contain activities that are often planned with preferential logic; maintenance projects would have multiple constraints and so on.

The size of the project influences as well some of the metrics thresholds. The higher the total installed cost is, the greater the schedule tends to be in terms of number of activities, the longer the project's lifecycle and the smaller the granularity. The notion of 'high duration' and 'high total float' doesn't have the same meaning for a 6-months project or for a 10-year project. Whether or not the project schedule was developed on a rolling wave methodology also needs to be considered. Just as with the specification of percentages of activities meeting a stated metric, high duration activities should be defined as a percentage of the project length.

Amongst the most popular checks included in the diagnostic protocols and listed above, the disagreement between metrics persists on the thresholds to apply, as highlighted below:

1. Logic. This is one of the most frequent metrics used in a schedule diagnostic protocol. Any scheduler is familiar with the mantra "every activity needs at least one predecessor and one successor, with the exception of the first and last milestones in the chain". Any CPM schedule belonging to a Level 3 or higher should be subjected to this check. While most protocols seem to accept a 5% threshold for the total number of activities that are missing a predecessor or successor, any result higher than 2 activities with missing logic should at least trigger a review. A single activity that should belong in the critical path and is not integrated in the network can lead to a schedule with a fatal flaw. The deviation to the rule can be accepted on a case-by-case basis, but relying on the 5% rule is a risky practice;

2. Leads (negative lags). With thresholds varying from zero, to less than 1%, to less than 5%, to “not specified”, this is one of the contentious metrics. However, most guidelines seem to adhere to the view that, as long as the relationship models the way the work is to be executed, the use of a lead in a schedule is acceptable. In most cases the value of the limit, while subjective, is meant to trigger a further schedule review and not a schedule rejection;
3. Lags. Most schedule checks seem to concur on the “less than 5%” threshold. This limit is questionable and depends largely on the schedule class and granularity. For example, for class 3 schedules, lags are expected, as the schedule is developed at a higher level. This view is reflected in the AACE Recommended Practice 27R-03, Schedule Classification System that states ‘The Class 3 schedule […] should be developed using relationships that support the overall true representation of the execution of the project (with respect to start to start and finish to finish relationships with lags)’ [2, p. 8]. The definition asserts that start-to-start and finish-to-finish relationships with lags are useful and endorsed, when appropriate;
4. Relationship types. The DCMA 14-point protocol and Oracle’s P6 schedule check use this metric, with a not less than 90% threshold set for the finish-to-start relationships. No other standard or guideline endorses this metric, that would be difficult to defend. As mentioned above, schedule classes 3 and higher are expected to contain start-to start and finish-to-finish relationships with lags;
5. Hard constraints. The use of this term was not industry-standard when DCMA introduced it in its 14-point protocol. Hard constraints is now commonly used to describe the ‘two-way’ constraints where an activity is fixed to a certain date regardless of logic. As with the ‘Leads’ metric, the thresholds proposed vary based on the standard or tool used, from zero, to less than 1%, to less than 5%. For Schedule Classes 3 or lower and irrespective of the lenience of the standard, any hard constraint needs to be properly documented and justified;
6. Soft constraints. Not often included in the schedule checks, this metric is sometimes combined with the ‘hard constraints’ one and its limit vary between less than 5% to less than 10%. This is another one of the metrics that requires further evaluation. Some soft constraints are harmless to the schedule integrity. For example, in detailed execution schedules, contractor mobilization activities or required on site (ROS) milestones can be assigned as late as possible constraints with positive lags, to ensure that they remain aligned with the construction requirements. From a metrics analysis perspective, the schedule contains a multitude of milestones, relationships with positive lags and soft constraints and a possible Failure score for these metrics, that in reality do not compromise the schedule integrity.
7. Negative float. While proscribed by DCMA and Acumen Fuse Schedule Index, this occurrence seems to be acceptable in Oracle’s P6 Schedule Check, up to a 1% limit. While other standards do not specifically address the issue, negative float is valid in certain circumstances, such as change orders not yet processed. In any instance, this occurrence should have a stated justification.
8. High (large) activity durations. Often included in the schedule checks, this metric has default values of 44 working days or 60 calendar days, and a threshold of less than

5%. For Level 3 schedules, especially for large and mega projects, both the duration value and the threshold usually need to be adjusted. Acumen Fuse's 'Insufficient detail' metric is replacing the fix default 44 days duration with a variable equal to 10% of the project's duration, partially addressing the scalability issue. Fabrication activities or high-level engineering activities used only for resource loading purpose would still need to be excluded from this check;

9. Logic Density. For this metric, Deltek sets the threshold to minimum two and maximum four relationships per activity, with a higher number indicating an 'overly complex logic'. While the inferior limit of two relationships per activity is valid, users would have difficulty understanding why more than four relationships would be bad. As an example, in EPCM Level 3 schedules, for the installation activities, the following prerequisites would be required:

   a. the contractor to have been mobilized
   b. the equipment to have been delivered
   c. the material supplied by contractor to have been delivered
   d. the drawings to have been issued to the contractor (unless all drawings were issued at contract award)
   e. the access to the area to have been granted (especially if provided in stages)
   f. the physical constraints (steel erection after foundation, cladding after steel erection etc.)

   and this without including any preferential logic or resource driven logic.
   While it is true that not all these relationships would be required for all the installation activities, chances are high that the upper limit of four relationships will be exceeded. In comparison, NAVAIR's schedule check sets the upper limit to ten relationships for its equivalent metric.

As for the proprietary metrics, Deltek's Merge Hotspot metric does not have an equivalent in the industry standards. As with the Logic Density metric above, its application on schedules Levels 3 and lower is not appropriate, but even for Class 2 and Class 1 schedules, developed to a Level 4 granularity, the ideal target of less than 10% activities with no more than 2 predecessors is difficult to defend. On the opposite side, it can be argued that schedules that meet these criteria are under-developed and have insufficient logic.

Among other checks performed by the software providers, some are very helpful and could indeed reveal structural issues within the schedule model. As an example, the Open Ends metric helps identify the dangling activities. However, in the Acumen Fuse application, this analysis incorrectly flags milestones as having open ends, and leads to start milestones linked start-to-start to its successor or finish milestones linked finish-to-finish to its predecessor to appear as having an open finish or an open start, respectively. For an accurate analysis, these occurrences would need to be omitted from the check.

Other metrics are just informative, in that they give additional insight into the schedule structure, without a predefined threshold. For instance, minimum lag, maximum lag, number of activities, either normal tasks, level of effort or summary tasks, number of constraints, by individual type, can add another layer of information into the schedule diagnostic. It is up to schedule reviewers to look into these metrics further if particular parameters seem problematic.

Using the Appropriate Evaluation Protocol

The evaluation protocols to be used in any schedule investigation differ whether or not the schedule under evaluation is a baseline schedule or an update schedule. Baseline schedule reviews should concern themselves with quality and completeness issues; the metrics used to evaluate them convey mainly information on the schedule's structural integrity. Update schedule reviews should primarily look at status and schedule changes [9] [10]. For update schedules, metrics such as missed tasks, baseline execution index (BEI), invalid dates, 'Hit or Miss' ratio, are a primary focus, providing additional information into the project's adherence to the baseline plan.

As baseline schedules usually do not contain actual dates [9, p. 8], their evaluation concerns all the activities in the network. Some activities might subsequently be excluded from the analysis, based on the metrics that are evaluated, as to depict the schedule structure more adequately. A first selection can be made by activity type, filtering out level of effort or summary activities for example. Other selections can be made to zoom in specific groups of activities, using the same type of filters as when developing layouts in the scheduling software.

For update schedules, although their structural soundness is expected, the metrics evaluation usually focuses on the activities that are not completed. This view is shared by most standards and guidelines. However, Deltek adopts a different approach and, by default, in the calculation of the Schedule Quality Index, with the exception of Critical and Negative Float metrics, all other metrics are counting the completed activities as well. As with the other customizations that are possible within the tool, the selection of the activities to include in the analysis is then left at each user's discretion.

But within any one category, either for baseline schedule review or update schedule review, the protocols might differ in implementation and lead to different results. In the case of the 'in-house' tools developed by users in simple spreadsheets, there is a risk of errors originating from incorrect formulas or data corruption. When a software solution is used for the evaluation, the different results obtained could be attributed to the version of the standard used by the software, or by the interpretation of the standard requirements by the developers of the respective protocol.

To prove the point, a Class 3, Level 3 schedule, prepared at funding request, was analyzed based on the DCMA-14 protocol, using two different software: Acumen Fuse and Schedule Analyzer. This protocol was selected simply because it is common to both software providers. It is important to note that the submitted schedule was not actually subjected to the DCMA-14 protocol. The results are included in Table 6 below.

Nb.	Metric	Fuse Result	Schedule Analyzer Result
1a	Logic / Missing Logic	13 (2%)	82 activities Ratio = 1.51%
1b	Dangling Logic		82 activities
2	Leads	0 (0%)	0
3a	Lags	451 (30%)	572
3b	Long Lags		569
4	SS/FF Relationship Count	479 (32%)	479
4	SF Relationship Count	0 (0%)	0
5	Hard Constraints	0 (0%)	0 or 0.00%
6	High Float / Total Float > 44d	447 (52%)	449 or 52.27%
7	Negative Total Float	0 (0%)	0
8	High Duration / Original Duration > 44d	238 (28%)	238 or 27.71%
9a	Invalid Forecast Dates	0 (0%)	0
9b	Invalid Actual Dates	N/A	0
10	Resources	470 (55%)	415
11	Missed Activities / Missed Tasks	0 (N/A)	N/A
12	Critical Path Test	(check)	Instructions for P6 Check
13	CPLI (Critical Path Length Index)	1.00	1
14	BEI (Baseline Execution Test)	N/A	N/A
	Protocol Version Used	Not indicated	Based upon "OMP/IMS Training" ENGR120 presentation dated 21-11-09

Table 6. Example DCMA-14 Point Analysis Results from Two Metrics Software

As indicated in Table 6 above, Acumen Fuse does not specifically state the DCMA-14 point protocol version used for the evaluation, while Schedule Analyzer bases its analysis upon ‘OMP/IMS Training’ ENGR120 presentation dated Nov 2009.

Without making a case for the use of one software over another, as seen above, for the same protocol, some metrics present different results: logic, lags and resources. This can prove confusing for users and require additional validations of the results.

The analysis above also highlights the importance of using the correct protocols for any given schedule. The sample schedule selected, prepared at funding request, was submitted to quality checks using a customized protocol, derived from the Acumen Fuse Schedule Index. The DCMA-14 protocol, while not requested for this schedule, proves to be inappropriate in this case. To note that the project for which this schedule was prepared was completed successfully, ahead of its target, and was also recipient of two major Canadian awards. This only to prove the point that not meeting a certain threshold, that is inadequate from start, has no negative consequence on the schedule integrity.

The implementation protocols also vary for the update schedules. Calculating some of the metrics, such as 'Missed activities', 'CPLI' and 'BEI', requires using baseline information. Knowing that the native .xer files do not contain baseline information, raises the question as to what fields Acumen Fuse reads when the schedule is prepared in P6 and then imported as .xer file. As it turns out, Acumen Fuse is considering the planned start and planned finish as baseline start and baseline finish. Expert P6 users know that this is not the case, so these metrics are simply incorrectly calculated. This author has recommended a product change on this basis.

Another schedule diagnostic tool, Schedule Analyzer, connects directly to the P6 database and does not encounter this issue. The user is specifically asked if the target (older) schedule is the currently approved baseline schedule (as per DCMA-14 point specification) and if the user indicates 'no', then the program requires the user to select the correct schedule before proceeding further [13].

Understanding the System Configuration

When deciding to use alternate evaluation protocols, developed by the various software providers, users need to understand the system configuration options. For example, the results of any Fuse analysis are dependent on the scoring option selected prior to the analysis [14] [15] and on the metrics customization.

By default, Acumen Fuse uses a record-based scoring method. This method requires activities (or 'records') to pass all metrics selected for the analysis and computes the final score based on a pass/fail criterion, counting the number of activities that passed all metrics versus the number of activities that failed any of the selected metrics. In the metrics editor, users can designate the selected metrics as ‘bad’, ‘neutral’ and ‘good’ and any occurrence of ‘bad’ metrics negatively influences the overall score.

The second evaluation option offered by Acumen Fuse, the metrics-based scoring, allows activities to receive partial credit towards the overall score, based on the number of metrics that pass the criteria and the weighting values attributed to those metrics. Deltek states that, by default, the metrics weighting values were set to the midpoint in the weighting scale, but this doesn’t always seem to be the case. Open ends, minimum lag, maximum lag, leads, negative float, more than 30 days float, are just a couple of metrics set to the absolute ‘bad’ values in Acumen Fuse metrics library [14].

The selected score calculation method greatly influences the overall score results. Using the same sample schedule as before, below are the results of the Acumen Fuse analysis for the overall Schedule Index, using each of the available scoring methods.

Metric	**Result**	**Result (%)**
Missing Logic	17	2%
Logic Density™	3.62	
Critical	169	15%
Hard Constraints	0	0%
Negative Float	0	0%
Insufficient Detail™	44	5%
Number of Lags	448	40%
Number of Leads	0	0%
Merge Hotspot	197	17%
Score based on Record-Based Method	**50%**	
Score based on Metrics-based Method	**94%**	

Table 7. Acumen Fuse Schedule Quality Analysis for Both Scoring Methods

While all individual metrics results are identical, the overall score might generate a pass or fail results, based on the selected scoring method.

Moreover, using the same scoring method, users can influence the overall score results by changing the weightings applied to the selected metrics. As an example, another analysis was performed under the record-based method option, modifying the weightings for the Number of Lags and Merge Hotspot metrics from -5 (bad) to 0 (neutral). The overall score changes from 50% to 95%.

While a very powerful feature due to the great flexibility it provides, all these customizations are not transparent to the readers when viewing the analysis report, and arguably are not known to every user of this software. When publishing the analysis report, the details behind the calculations are not included, and it is up to each user to document all the changes made for customization.

Schedule Benchmarking

As the concept of benchmarking the schedule quality starts to make way in the industry, the obvious question to be asked is, are the benchmarking data reliable? To answer this question, the very nature of assessing the schedule quality needs to be considered. What data was collected in any previous schedule evaluation, at which moment in the project's lifecycle, from what source, using what diagnostic protocol? What controls are put in place in a benchmarking exercise, to make sure that the comparison is made on similar bases? As the projects are, by definition, unique endeavors, any benchmarking data would need to be normalized, requiring a significant amount of information that documents the project's context.

Acumen Fuse introduced a benchmarking feature, intended to allow users to compare the results of their schedule quality assessment to the ones residing in a cloud-based metric library. This library contains sets of metrics collected from schedules developed by other Acumen Fuse users, presumably from the same industry. Using a simple protocol, the metrics from the schedule being evaluated are individually compared with the ones in the corresponding sets from the library, and the overall score for either Fuse Schedule Index or Fuse Logic Index is plotted on a graph [16]. Although it is beyond the scope of this paper to investigate the inner workings of this functionality, the authors note that, while visually attractive, this feature presents a number of flaws.

The information that comprises the benchmarking data lacks transparency. There is no available information as to the number of schedules that comprises the data set, the schedule class, the schedule level, the project development stage, the size of the project, the scheduling party (owner, contractor, EPCM company), and so on. All of these are factors that would need to be taken into account for a benchmarking exercise and for normalization purposes. Other than the category / sub-category criteria used to group projects by industry, there are no additional criteria to zoom-in for a specific set of data.

Once the software users opt-in for sharing their own set of data, the corresponding metrics are collected by the software company. The data set will then be associated with a specific project type, based on the category / sub-category selected by the user. If no category is assigned by the user, the project gets put into a pool of ‘uncategorized’ projects. If a user would want to compare his results with the ones in other industries, he would select a different category and launch the feature. Due to a current flaw in the product, the same data set could be collected for more than one category of projects, as each use of the functionality will trigger the sampling. An enhancement request was raised by this author and it is hoped that the data collection procedure will be modified in a future version of the software to correct this flaw.

On the same subject of data collection protocol, there is another weakness. Consider a schedule that is in the development stage. As the schedule matures, the schedulers could use Acumen Fuse to self-asses the quality of their schedule and use the results to bring it to adherence standards, whatever they might be. In this case, each iteration will get stored and the software does not allow for differentiation between a work in progress schedule and a final schedule. Moreover, as the schedule is getting ready to be issued for client review, another analysis could be performed by the functional team. Finally, as the schedule is sent over to the client for approval, another analysis can be made by the client, using the same software. And if iterations are required to fix outstanding issues, this will just multiply the number of instances that get collected. The data collection protocol would need controls in place to ensure no duplication of the same data set is stored.

From a practical point of view, the benchmarking feature allows the comparison to be made based on two different overall scores: Fuse Schedule Index and Fuse Logic Index. In each case, the default scoring method and weighting criteria are used for the comparison. During the software customization process, if the user decides to remove or to add metrics in the analysis, to modify the weightings or the thresholds as to suit its project characteristics, then all those customizations are ignored by the benchmarking feature. A schedule could fare very well on a custom quality index, but then would be ‘normalized’ and brought into the mold of the Fuse standard analysis. Currently, the benchmarking feature can’t be customized to include a specific set of metrics for comparison, and this could prove confusing for the report readers.

Other software providers are using a different version of the benchmarking feature, allowing users to create reports from their own scheduling database. One example of such providers, Schedule Analyzer Enterprise Forensic software package, contains a benchmark module that allows users to create benchmarking from their own scheduling database. Using the out-of-sequence report, lag analysis report, or activity duration and float report, planned versus actual scheduling statistics can be amassed and standard of deviations defined [17] [18] . This type of analysis works well to track project's performance during the execution phase.

As standards and guidelines covering the schedule quality evolve, the software providers follow suit and adapt their packages. With the rapid advancement of technologies, it is expected that even more schedule analytics products will appear on the market. But users need always to be wary as to how these products work and get involved in addressing the discovered flaws. Bugs identifications and enhancement requests are good ways of communicating to software providers technical issues. These processes are meant to improve the products usability and ultimately help the scheduling community.

Recommendations for Using the Appropriate Schedule Metrics

As seen above, a multitude of metrics can be used to analyze the structural integrity of the schedules during their development and throughout most of the project's life cycle. This metrics analysis is best used as a first step in a more comprehensive schedule review process. Its results will only attest if the schedule dates can be reasonably trusted, and nothing more. Subsequent analysis will treat scope completeness, accurate reflection of the execution strategy, the plan's feasibility, and the like. Performing these detailed analyses on a flawed schedule model risks being a waste of time.

But not all schedule statistics are equally important in measuring schedule quality. Some metrics are more critical than others, based on the specificity of the schedule that is being analyzed. Some thresholds are stricter than others, and trigger schedule revisions or even schedule rejections. And some metrics are informative only, and do not trigger any further action. Any schedule quality analysis should be concerned with focus, appropriate protocol, and normalization.

1. Focus. At the onset of any analysis, users need to ask a series of questions to better define the framework of the exercise. What is the objective of the analysis? Is the analysis performed as a self-assessment of schedule quality or is it performed to ensure compliance to a given standard? Is it a company standard or a client standard? The answer to this question would indicate the degree of flexibility with the results obtained and the actions to follow the analysis. What are the activities to retain in an analysis? Does the analysis cover the entire schedule or a specific portion of it? The answer will determine the filters to be used in the analysis and the data set that will be investigated.

2. Appropriate protocol. Before proceeding with the analysis, the appropriate protocol needs to be selected based on the schedule type, either baseline schedule or update schedule. Specific structural requirements apply to baseline schedules that might be acceptable in update schedules. For example, the test for negative total float is not acceptable in baseline schedules but might have an acceptable justification in update schedules. The metric for identifying missing logic, that could trigger corrections in baseline schedules, might prove less critical in update schedules, especially if the activities in question are completed activities. Moreover, as update schedules should be judged on the accuracy of status reporting and adherence to the work plan, additional metrics should be included in the analysis to track these parameters.
3. Normalization. Most of the evaluation protocols, including DCMA-14 points assessment, were designed specifically for detailed control schedules, usually developed to Level 4. When used in other circumstances, their utilization should be normalized for that level. For example, Level 3 schedules are developed to a lower granularity, requiring the use of relationships start-to-start and finish-to-finish with lags. High durations are also likely to occur more often in Level 3 schedules or in schedules adopting a rolling wave methodology. When left unchanged, these protocols are likely to generate fail results in acceptable schedules Level 3 and lower.

After adopting the appropriate protocol and normalizing the individual tests, the review analysis can be performed in simple spreadsheets, in-house developed tools, or commercial software. When selecting the latter, users would need to give special consideration to software customization, transparency, and results validation.

1. Software customization. Based on the software solution used for the analysis, users might have more or less flexibility in customizing the test by modifying the selection of metrics to be included in the analysis or their thresholds. The proposed schedule quality indices should be seen as guidelines only and they should prompt reflection. The users should strive to develop schedule quality indices that take into account the particularities of the schedule that needs to be evaluated. A good analysis should capture all relevant metrics for that specific schedule. This would include everything that would otherwise compromise the use of the schedule for the purpose it was developed for. Considering the multitude of metrics included in the protocols listed above and many more others available in the various software packages, some metrics are more critical than others at various points in the project lifecycle.
2. Transparency. Irrespective of the tool used for the analysis, users should be transparent in the methodology used, in the metrics selection criteria, and in the thresholds used. All assumptions, additional filtering of the information, and tool customizations, including the calibration of the metrics weighting, when applicable, should be clearly described in the schedule basis document.
3. Validate the results obtained. While software continues to evolve and new products enter the market, users need to keep their guard up and always validate and question the outcomes they produce. Just because a software issues a certain result, this doesn't necessarily make it 'right'. By asking questions and raising flags on software

abnormalities, the scheduling community can help software companies address them, for the common good.

Conclusion

The metrics evaluation seeks to determine if the schedule model is correctly constructed, so as to provide reliable information on the calculated dates. The difficulty of this approach is in the implementation of a metrics analysis.

With reference to the original question, the review of the current public schedule metrics shows a widespread confusion as to which metrics to use. Guidelines such as the DCMA 14-Point Assessment, NASA's Scheduling Management Criteria, NAVAIR 11-Point Assessment, NDIA were introduced. Schedule metrics software implementations such as Oracle's P6 14-point checks, Acumen Fuse Schedule Index, Schedule Analyzer for the Enterprise's Baseline Checker, and SmartPM Schedule Quality Index were also briefly introduced.

General issues with using the correct metrics were presented. Schedule class and level has a definite bearing on metric appropriateness. The size of the project being investigated definitely influences the thresholds being used to assess schedule quality. Each schedule quality analysis should be concerned with focus, appropriate protocol, and normalization, and it was highlighted that software implementations require proper customization, transparency, and validation of the results obtained.

The bodies that issue standards unfortunately do not test and certify software implementations. This can lead to various versions of the same guideline being used by different software providers, or to different interpretations of a given metric. Many of the software products proport to measure the same thing but can deliver different results Moreover, different results can occur using the same software, but with different 'standard' settings. How the proprietary metrics are being used has a direct bearing on the evaluation of a metric's conclusions.

Developing a schedule is both a science and an art. Enforcing adherence to certain metrics and inappropriate thresholds reduces a scheduler's flexibility in the schedule development process and does not necessarily ensure a better schedule outcome. Some schedulers are required to use DCMA 14-Point Assessment due to contractual obligations with the DOD or other project owners who adopt this out of expediency, but this is not the case with other proposed schedule quality indices.

The users voluntarily using scheduling metrics regimens need to make an informed decision into the criteria used to determine their schedule quality. Used mindfully, selecting the right metrics, the right criteria, the right limits for each particular schedule, could help practitioners deliver technically sounder schedules. This requires renouncing the ‘one size fits all’ current implementation trend of these diagnostic tests.

And while benchmarking is a sought-after undertaking, the tools available on the market do not provide yet a suitable solution. Acumen Fuse’s benchmarking feature, while interesting in concept, requires additional controls in place to ensure reliable data collection and normalization before a wide-spread adoption of the tool can be envisioned. Other software tools for benchmarking various schedule statistics should also be considered and evaluated.

In conclusion the question remains, can the metrics proposed by the various software providers be relied upon without an industry validation? Innovation is welcome but should be promoted with transparence and support data. Just because a software provider includes a certain metric in its protocol or proposes a default threshold for a given metric, this does not make them an ‘industry standard’. System implantation options used should be thoroughly understood and documented and users should take an active role in helping software companies fix the identified bugs.

References

[1] Ronald M. Winter, "DCMA 14-Point Schedule Assessment," AACE International Transactions, Morgantown, WV, 2011.

[2] AACE International, "Recommended Practice No. 27R-03 - Schedule classification system," AACE International, Morgantown, 2010.

[3] Defense Contract Management Agency, "DCMA-EA PAM 200.1, Earned Value Management System (EVMS) Program Analysis Pamphlet (PAP)," Department of Defense, Washington, DC, 2012.

[4] NASA, "NASA Schedule Management Handbook, NASA/SP-2010-3403," NASA, Jan 2010.

[5] U.S. General Accountability Office, "Schedule Assessment Guide Best Practices for Project Schedules (GAO-16-89G), First Edition," U.S. General Accountability Office, 2015.

[6] NAVAIR, "Integrated Master Schedule (IMS) Guidebook," Public Release 09-456, Feb 2010.

[7] National Defense Industrial Association, "Planning and Scheduling Excellence Guide (PASEG), Version 4.0," National Defense Industrial Association, September 2019.

[8] D. Patterson, "The Value of a Standard Schedule Quality Index," 2012.

[9] AACE International, "Recommended Practice No. 78R-13, Original Baseline Schedule Review – As Applied in Engineering, Procurement and Construction," AACE International, Morgantown, Latest revision.

[10] AACE International, "Recommended Practice No. 53R-06, Schedule Update Review - As Applied in Engineering, Procurement, and Construction," AACE International, Morgantown, Latest revision.

[11] SmartPM, "Understanding the Schedule Quality Index," SmartPM, June 2019.

[12] C. Shoshanna Fraizinger, "Which Schedule Quality Assessment Metrics to Use? … and When?," AACE International Transactions, Morgantown, WV, 2019.

[13] "Schedule Analyzer for the Enterprise (SAe) Version 4 manual," August 2017.

[14] "Deltek Acumen 8.5 Online Help," Latest revision.

[15] M. Wiechec, "Applying and understanding the Acumen Fuse Schedule Quality Index™," HKA Global Ltd., 2017.

[16] Ten Six, "Deltek Acumen Benchmarking Feature," Ten Six, July 2020.

[17] "Schedule Analyzer Enterprise Forensic (eForensic) Version 2 manual," October 2020.

[18] Ronald M. Winter, "Mastering Out-of-Sequence Progress – Part 3," AACE International Transactions, Morgantown, WV, 2019.