Journal of Cost Analysis and Parametrics, 7:160–179, 2014
Copyright © ICEAA
ISSN: 1941-658X print / 2160-4746 online
DOI: 10.1080/1941658X.2014.982232

www.iceaaonline.com

An Assumptions-Based Framework for TRL-Based Cost and Schedule Models

BERNARD EL-KHOURY[1] and C. ROBERT KENLEY[2]

[1]The Boston Consulting Group, Dubai, United Arab Emirates
[2]Purdue University, West Lafayette, Indiana

The Technology Readiness Level scale has been used to assess progress and provide a framework for developing technology. Many Technology Readiness Level-based cost and schedule models have been developed to monitor technology maturation, mitigate program risk, characterize transition times, or model schedule and cost risk for individual technologies as well technology systems and portfolios. We present a four-level classification of models based on the often-implicit assumptions they make. For each level, we clarify the assumptions made, review evidence that supports the assumptions, and propose alternative or improved models. Our results include a justification of the recommendations of the US General Accounting Office on Technology Readiness Level, two new methodologies for robust estimation of median transition times and for forecasting transition times using historical data, and a set of recommendations for Technology Readiness Level-based regression models.

Introduction

Innovation and technology are key elements to maintaining a competitive edge and sustaining growth. Companies as well as state agencies must develop new technologies that are timely and cost-effective and that reduce the risk of schedule and cost overruns. Technology development is highly unpredictable: Not only is a project manager faced with "known unknowns" (i.e., uncertainties that can be roughly estimated and controlled), the manager also has to deal with "unknown unknowns," which are completely unforeseeable uncertainties due to the very nature of developing a new technology. Industry has adopted many technology management frameworks to address this challenge, such as technology roadmapping, technology benchmarking, technology watches, and technology risk management (Foden & Berends, 2010). The aim of these frameworks is to control key factors such as cost, schedule, technology maturity, and manufacturability.

The US Government Accountability Office recommended, "Maturing new technology before it is included on a product is perhaps the most important determinant of the success of the eventual product—or weapon system," and GAO encouraged the use of "a disciplined and knowledge-based approach of assessing technology maturity, such as TRLs, DoD-wide" (GAO, 1999). Technology Readiness Levels (TRLs) are a 1-to-9 scale, as shown in Table 1, developed by NASA (Mankins, 1995) that describes the maturity of a technology with respect to a particular use. The scale captures the maturity of the technology by looking at major milestones in the technology development process and is intended to be consistent across technologies. GAO suggested the use of TRLs to make

Address correspondence to C. Robert Kenley, Associate Professor of Engineering Practice, 315 N Grant St., West Lafayette, IN 47907-2023. E-mail: kenley@purdue.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ucap.

TABLE 1 NASA TRL scale definitions

TRL	NASA TRL Definition
1	Basic principles observed and reported
2	Technology concept and/or application formulated
3	Analytical and experimental critical function and/or characteristic proof of concept
4	Component and/or breadboard validation in laboratory environment
5	Component and/or breadboard validation in relevant environment
6	System/subsystem model or prototype demonstration in a relevant environment (ground or space)
7	System prototype demonstration in a space environment
8	Actual system completed and "flight qualified" through test and demonstration (ground or space)
9	Actual system "flight proven" through successful mission operations

sure that technologies are mature enough before integrating them into systems. Within Department of Defense (DoD), organizations that perform research and development use TRLs to understand a technology's riskiness by looking at its maturity (Graettinger, Garcia, & Ferguson, 2003), and organizations that acquire systems use TRLs in Technology Readiness Assessments (TRAs) to make sure the technology has matured enough to pass certain milestones (DoD TRA Deskbook, 2009).

Azizian, Sarkani, and Mazzuchi (2009), Cornford and Sarsfield (2004), Nolte (2008), and Fernandez (2010) pointed out that TRL is not well integrated into cost, schedule, and risk modeling tools. Many models have been proposed to use TRLs for cost and schedule modeling for individual technologies (e.g., Conrow, 2011; Dubos, Saleh, & Braun, 2008; GAO, 1999; Smoker & Smith, 2007) as well as technology systems and portfolios (e.g., Dubos & Saleh, 2011; Lee & Thomas, 2003; Sauser, Ramirez-Marquez, Magnaye, & Tan, 2008). These models are based on different assumptions that can lead to different results.

We have developed a framework to explain these differences. When evaluating a model from the literature on approaches of using TRL to model cost and schedule risk, we often identified another model that had more useful results, but at the price of stronger (often unstated) assumptions. We have grouped the into four levels of increasingly strong assumptions, which facilitate comparing, developing, and improving TRL-based cost and schedule models using the most relevant characteristic: the assumptions that they make.

Figure 1 shows our unifying framework for models that use TRL based on the assumptions that they make. For each level in the framework, we will (1) state the assumption, (2) review the literature relevant to that level of assumption, (3) describe evidence supporting this assumption, and finally, we will (4) propose new methodologies that make better use of the assumptions whenever possible. The level-1 assumption that TRL is a time-ordered scale is relatively weak, whereas the level-4 assumption that TRL is a metric scale is the most strong. Lower levels in our framework permit simple, but robust models, and higher levels permit more advanced and more useful models, but that are less robust due to the strong assumptions.

At level 1, we will answer the question:

Does available evidence support or contradict GAO's recommendation on pre-production maturation?

At level 2, we will answer the questions:

Does it make statistical sense to look at the distribution of each TRL transition time separately?

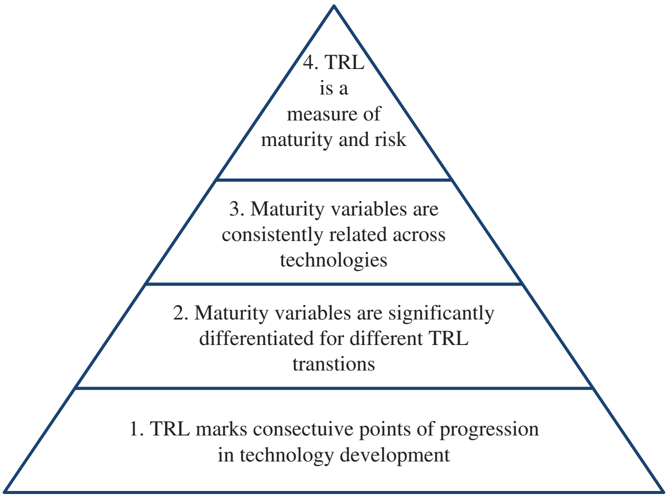

4. TRL is a measure of maturity and risk
3. Maturity variables are consistently related across technologies
2. Maturity variables are significantly differentiated for different TRL transtions
1. TRL marks consectuive points of progression in technology development

FIGURE 1 The four-level framework for TRL-based cost and schedule models.

If yes, is there a way to improve the accuracy and fidelity of the estimates and confidence intervals?

At level 3, we will develop a model that answers:

If we use historical data of a technology's development, can we significantly improve the accuracy of the technology schedule forecast over level-2 models?

And at level 4, we will answer:

Out of the available methodologies, what is the best one to use in performing regression of cost and schedule against TRL?

Transition Variables and Schedule Transition Datasets

We define a "transition variable" as a management-relevant variable that is associated with the transition between two TRLs. The schedule transition variables will be noted $X_{i\text{–}j}$ and the cost transition variables $C_{i\text{–}j}$ (both are considered random variables). For example, $X_{3\text{–}4}$ corresponds to the time the technology takes to transition from TRL 3 to TRL 4, and $C_{3\text{–}8}$ is the cost incurred in transitioning the technology from TRL 3 to TRL 8.

We used two schedule transition datasets: a NASA dataset is a relatively high quality dataset, but it has only 19 technologies; and an Army dataset is larger (582 technologies), but with lower quality.

The NASA Dataset

The first dataset is from a case study done by the Systems Analysis Branch at NASA's Langley research center which looked at typical times that aeronautical technologies take to mature (Peisen, Schulz, Golaszewski, Ballard, & Smith, 1999). They collected the data through interviews with NASA personnel. Table 2 shows the full data set. The columns constitute different NASA technologies that were selected for the study. The first eight rows correspond to the TRL transitions (the values of the variables are the durations in years of the technology TRL transition times).

The dataset does suffer some drawbacks:

1. There are a rather small number of technologies (only 19), and seven of those data points have only the first five transitions.

TABLE 2 The NASA dataset

Transition	Carbon-6 Thermal Barrier	Direct To	Fiber Preform Seal	Low Emissions combustors	Nondestructive Evaluation	Tailless Fighter	Thrust Vectoring Nozzle	Electro Expulsive DeIcing	Engine Monitoring Systems	Flow Visualization	Fly-by-Light	GA Wing	Graphite Fiber Stator Vane Bushings (Tribology)	Particulate Imaging Velocimetry	Propfan development	Runway Grooves	Surface Movement Advisor	Supercritical	Tiltrotor Technology
12	0.4	0.2	1	1	0.5	3	0.3	0.5	0.5	5	2.5	0.5	1.9	2	2.5	0.7	0.8	1.5	3
23	0.4	0.1	1.5	1	1	1	0.3	0.5	0.5	1	5	0.5	1.9	4	1	0.2	0.3	1	1
34	0.4	0.1	1.5	1	1	1	0.4	1	0.5	3.5	7.5	0.5	1.9	2.5	1.5	0.2	0.3	1	1
45	0.5	1.1	1.5	2	1	1	2	1	0.5	1	4	3	1.9	3	2.5	0.2	0.35	1	1
56	0.2	0.1	6	4	1	2	2	1	1	0.5	1.5	0.5	1.9	0.5	1	0.2	0.35	1	22
67								6	0	0.5	1.5	1.5	1.9	0.8	2.5	1	0	1	8
78								0.5	5	0.5	1.5	3	1.9	0.3	6	1	1.2	12	0
89								5.5	0	0.5	1.5	4	1.9	0.3	1	1	0.1	1	11

2. The sample suffers from selection bias since it only contains the technologies that made it to TRL 9 or 6. Many projects may be abandoned at early TRLs. The negative effect of this phenomenon is that it reduces the representativeness of our work so that it cannot forecast program termination risk and is limited to programs that make it to TRL 9 or 6. Its positive effect however, is that it eliminates one source of uncertainty, so that we will not have to worry about modeling highly unpredictable external variables (such as budget cuts, program termination, requirement changes, etc.) that can lead a program to stop.
3. The data comes from retrospective interviews based on descriptions of TRLs (as opposed to rigorous TRL calculators), which means it could contain measurement errors. Furthermore, the authors say that they had to do some smoothing when the interviewees did not remember the exact transition time; and that "some" points were at a one-year resolution level but did not indicate which points were at this resolution.

Inspection of the data reveals one clear outlier: Titltrotor Technology (which is the technology used in the V22 osprey) took 22 years to transition from TRL 5 to TRL 6, which is significantly larger than the other transitions in the dataset. Further inspection of the data reveals a positive skewness of the data: While most values are very small, a few transitions take relatively larger times, thus clustering the bulk of the data around zero. A log-transformation is natural in such cases of positive data with clear positive skewness. Three transitions had a duration of zero, and we performed a simple smoothing by assuming that this transition happened so quickly because of extra development effort in the preceding transition. We changed the preceding step to 80% of its original value and gave the remaining 20% to the "zero" step. We had two justifications for this modification. A theoretical one—a phase cannot be finished instantaneously—and a practical one—it is necessary to eliminate zero to perform a log-transformation on the data.

The Army Dataset

The Army dataset is larger than the NASA dataset, but of lesser quality. It contains partial TRL transition times of 582 technologies that were commissioned between FY 2005 and FY 2009. We extracted the data from Army Acquisition Technology Objectives briefings, which contained charts with the projected milestone and TRL dates for the critical components.

As a result, the dataset has two major quality problems:

1. It is based on contracts (promises) and projected schedules (as opposed to historical results). The final cost and schedule in DoD acquisition programs is often different from the initial estimates. Hence, instead of having a set of technology transition times, we only have a set of initial estimates of transition times.
2. Since the data was recorded visually from low-resolution graphs (values rounded to quarters of the year), it also suffers from low precision. Given the manner in which the TRL steps were extracted from the graphs, the error can be as high as 0.5 years.

The dataset also suffers from a minor issue. Although it contains transition information for a high number of technologies, the dataset does not contain a large number of transitions for a single technology. The dataset mostly contains one or two transition variables per technology (typically transitions from TRL 3 to TRL 6). This makes it harder to use the data to look for trends in the development of a single technology; however, it did allow us to evaluate the distribution of specific TRL transitions across technologies.

Analysis of the Level-1 Assumption

This first-level assumption is that "TRL marks consecutive points of progression in technology development" (Figure 1). While this assumption might sound very basic and trivial, it has important managerial implications since it directly relates to a recommendation by GAO (1999) on technology transition risk. This influential report advocating for the use of TRLs is the major example in the literature using this assumption. Furthermore, GAO recommended in the report that the maturing of technologies to at least TRL 7 to reduce risks and unknowns adequately before proceeding with engineering and manufacturing development. Other sources that specifically mention this basic reverse relation between TRL scores and technology risk include Legresley et al. (2000) and Nolte (2008).

This assumption comes from two basic properties of the TRL scale: The TRL scale is "complete" and "monotonic." The scale is "complete" when it covers the entire technology development space: Every technology has to be at a defined TRL at any point in time; and, throughout its development, it has to go through all of the 8 TRL transitions. The scale is "monotonic" when TRL always goes through those transitions in the same increasing order. When these properties are present, TRL will always mark consecutive points of progression (in an increasing order) in technology development until the TRL reaches 9 (or until the level at which the technology stops maturing as in the case of development is halted prior to TRL 9).

GAO (1999) described risk levels based on the assumption that the higher the TRL, the smaller the remaining overall uncertainty in managerial variables such as cost, schedule, and performance. A project at TRL 2 is subject to risks (cost, schedule, performance) on all the transitions from TRL 2 to TRL 9, while a project at TRL 6 is only subject to risks on transitions from TRL 6 to TRL 9. This is a direct consequence of the fact that all technologies have to go through all TRL transitions (completeness), and they have to go through them in order (monotonicity). In particular, the GAO report identified the TRL 6–7 transition as the threshold for going from high risk to low risk.

Using the NASA dataset, it is possible to quantify this uncertainty using the single maturity variable of schedule. We use the standard deviation of the "time to maturity" as a proxy for uncertainty, or the remaining schedule risk. In Figure 2, we plotted the standard deviation of the time to maturity (i.e., the remaining development time) for each TRL. For example the value indicated at TRL 3 is the standard deviation of $X_{3\text{–}9}$ (the transition time from TRL 3 to TRL 9), while the value indicated at TRL 4 is the is the standard deviation of $X_{4\text{–}9}$ (the transition time from TRL 4 to TRL 9), and so on. We can see that the time-to-maturity risk has very high values (standard deviation larger than 10 years) up until TRL 5. Once the project passes TRL 6, the standard deviation drops sharply to 5.6 years. This risk continues dropping and supports the GAO-defined threshold of low risk as that beyond TRL 7, because the transition to TRL 7 corresponds to a significant drop in schedule risk when compared to earlier stages.

Analysis of the Level-2 Assumption

The level-2 assumption states, "Maturity variables are significantly differentiated for different TRL transitions." This means that when we look at one technology transition, the maturity variables have a probability distribution different from other TRL transitions with a bounded variance. For instance, Peisen et al. (1999) noted that there is "considerable variability" in the time that technologies take to mature. The level-2 assumption stipulates that this variability is low enough to perform statistical forecasting of the technology maturation time.

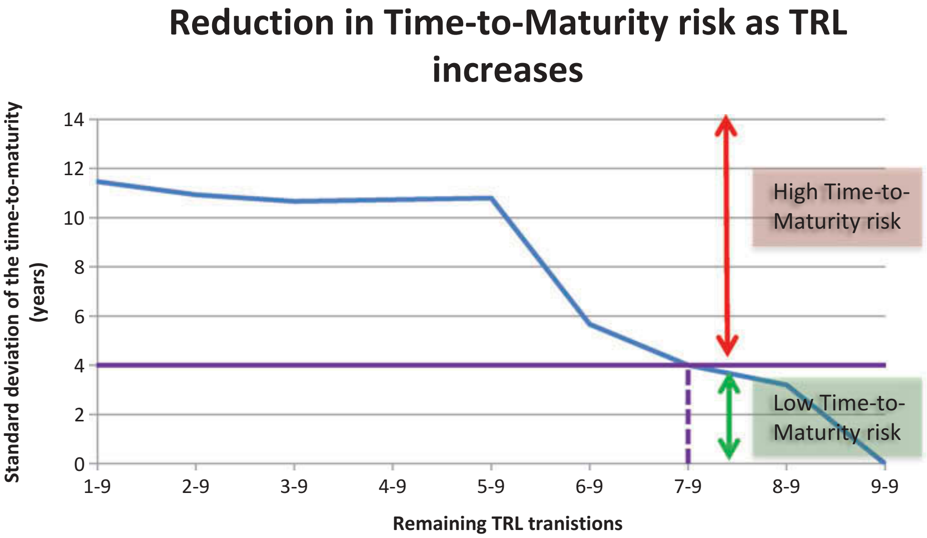


FIGURE 2 Reduction in time-to-maturity risk as TRL increases.

Many papers in the literature fall within this category of models, starting with Peisen et al. (1999). The authors characterized technology transition times by analyzing their distributions and by comparing averages and standard deviations of different subgroups of the sample. Similarly, Dubos and Saleh (2011) evaluated the distribution of every TRL transition time. They found that TRL transition times have lognormal distributions and used that to propose average estimators and confidence intervals. Finally, Lee and Thomas (2003) analyzed the distributions of absolute and relative cost growth at each TRL, and fitted the Johnson's distributions to the data (Johnson, 1949).

The very design of the TRL scale supports the level-2 assumption, because each transition in the scale corresponds to a well-defined action common to technology development. We should expect each of those transitions to share common properties across technologies, and thus we should expect each transition to be significantly differentiated from other transitions. For example, the TRL 1–2 transition that happens when "an application of the basic principles is found" is different from the 2–3 transition that corresponds to "going from paper to lab experiments," which itself is different from the 6–7 transition that happens when "the prototype is tested in the real environment."

Analysis of Variance for the Army Dataset

To validate level-2 assumption empirically, we looked at the distribution of different technology transition variables and tested if the transition times are precise enough to be statistically distinguishable. For example, transitions X_{1-2}, X_{2-3} are "distinguishable" if their variances are small enough so that they are both different from transition X_{1-3} using an Analysis of Variance (ANOVA). Otherwise, TRL 2 would be an unnecessary step, because introducing it would not add any statistically significant information over just using TRLs 1 and 3. In statistical terms, we are testing if the means of X_{1-2} and X_{2-3} are equal to that of X_{1-3} (Albright, Winston, & Zappe, 2006).

El-Khoury (2012) performed this statistical test for all pairs of TRL transitions where data was available using the Army dataset. The Army dataset is low resolution (low precision), but this does not fundamentally alter the results because the same imprecision in X_{1-2} and X_{2-3} are transferred to the sum in X_{1-3}. The imprecision gets cancelled out later when X_{1-2} and X_{2-3} are subtracted from X_{1-3} to compare the difference to zero in the test. The Army data consists mainly of promises as opposed to historical values, and we know that promises are biased and actual transition times will have higher variances than the ones agreed upon in the contract. To counter this problem, El-Khoury added random normal noise to the transition times in line with historical observations reported on DoD projects. El-Khoury performed eight statistical comparisons and equality hypotheses are rejected with very low p-values (<0.0001), indicating that the transition times are indeed significantly differentiated.

These results are beyond the ordinality property that we used in the level-1 assumption and show that the distance (in terms of maturation time) between the TRLs is statistically significant. This justifies the TRL definitions with respect to the "resolution" of the scale: Not only do the TRLs convey information about order on the scale, their definitions divide the scale in a representative manner such that we would lose information and precision about managerial variables such as schedule and cost if a TRL step were omitted from the scale.

Analysis of the NASA Dataset

Level-2 models consist mainly of building an empirical distribution of a transition variable and then using it to do estimations of important statistics.

The NASA dataset is less than 30 in size, and it is skewed, which also means that the population does not follow a normal distribution. As a result, we would have to make strong parametric assumptions about the distribution of the transition variable, assigning it to a specific known distribution with fixed parameters (e.g., "TRL transition time X_{4-6} follows a lognormal distribution with mean μ and standard deviation σ). Since TRL data is typically scarce, and because of its high skewness, we evaluated a more robust non-parametric measure such as the median for the average (we selected the median over the average because the latter is very sensitive to outliers and skewness, especially in small datasets).

El-Khoury (2102) evaluated three methods of non-parametrically generating median confidence intervals. First, Conover (1999) recommended using an approach for small datasets that is based on the binomial distribution. It ranks the values in the sample and then uses a table of values that he calculated to determine the confidence interval. Using the NASA dataset of 19 technologies, the 95% median confidence interval would simply be defined by the fifth value and the fifteenth one. Second, Olive (2005) proposed another technique as a modification of large-sample confidence interval formulation to improve the performance of the interval for small samples. Third, the bootstrap was recommended by Mooney and Duval (1993) especially in cases of asymmetrical data, even more so when the data is truncated and skewed (which is the case for all maturity variables). Figure 3 compares the 95% median confidence intervals generated using each of the three non-parametric methods. We can see clearly that the bootstrap confidence intervals (area highlighted in green) are the narrowest (except for the last transition where the CI generated by Olive's method was unrealistically reduced to the number zero, because this value appears many times in the X_{8-9} sample).

The bootstrap analysis presents some results that require explanation for those not well versed in statistical methods: For example, a 95% confidence interval could be the same as a 50% confidence interval. One way of overcoming this discrepancy in shape is by using the smoothed bootstrap described by De Angelis and Young (1992) that eliminates the discreteness of the median leading to smoother distributions that a project manager would be more familiar with.

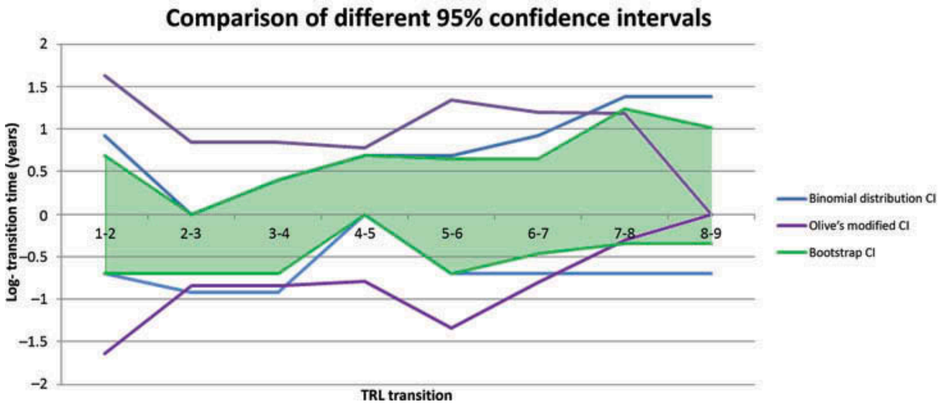


FIGURE 3 Comparison of 95% confidence intervals (of the median of log-transition time) generated using three different techniques.

Comparison of DOE, NASA, and Army Datasets

Crépin, El-Khoury, and Kenley (2012) analyzed Technology Maturity (TM) scores from a dataset from DoE's Nuclear Materials Stabilization Task Group. They transformed TM scores into TRLs and used multiple statistical tests to show that DoE TRL transition times were similar to NASA's TRL transition times (for example, $X_{4\text{–}5}$ values for NASA are not statistically different from DoE's $X_{4\text{–}5}$ values). This result generalizes the level-2 assumption across technologies and agencies.

Summary of Analysis of the Level-2 Assumption

In summary, the level-2 assumption is made whenever we want to study the statistics of maturity variables on each of the TRL transitions. For the Army dataset, an ANOVA supported the assumption that the TRL transition variables are significantly differentiated. While classical parametric estimators could be used, we recommend using the median bootstrap as a robust and accurate estimator better suited to small and skewed datasets (the datasets are small either from the start, or after undergoing segmentation for variance reduction).

Analysis of the Level-3 Assumption

The level-3 assumption states, "Maturity variables are consistently related across technologies." This level builds on the previous one: Not only are maturity variables well defined and differentiated, they are also related in a consistent (predictable) manner. It is clearly a stronger assumption because it means that if we pick any technology, we should expect some form of stable relation between separate transitions in the same technology. For management purposes, this means that for any project, we can use early transition information to make predictions on later transition variables (while level-2 models do not use any past information).

This assumption is made whenever we make statements such as: (1) "If a technology is already maturing quickly or cheaply relative to the average, then it is more likely to continue doing so for the remainder of its development," or (2) "High research and development costs are related to high development and engineering costs."

Table 3 shows the results of a correlation analysis of the NASA dataset and highlights correlations larger than 0.5. We can clearly see a cluster of highly correlated transition times 1–2, 2–3, 3–4, and 4–5. The rest of the transitions do not show any correlations except for a (minor) positive correlation between 6–7 and 8–9.

In simple terms, this initial cluster of positive correlation means that if a technology is maturing fast (resp. slow) in early TRL stages, then it is likely to keep maturing fast (resp. slow) in later stages. This phenomenon is true up to TRL 5. For later TRLs, however, we can see that transitions 6–7, 7–8, and 8–9 are very uncorrelated with all the early TRL transition times, as if they were independent of them. Perhaps this is because a technology changes hands after TRL 6, going from performing research in a laboratory to developing an operational system (Peisen Schulz, Golaszewski, Ballard, & Smith, 1999).

El-Khoury (2012) developed several forecasting methods and compared their performance. The forecasting methods by category are as follows:

Fixed estimates
- Mean
- Median

Influence Diagrams
- ID (full)
- ID (frag 4–3)
- ID (frag 5–2)
- ID (bounded)

Extrapolation
- Moving average
- Exponential smoothing
- Exponential smoothing (with trend)

Regression
- Full Autoregression
- Full Autoregression (bounded)

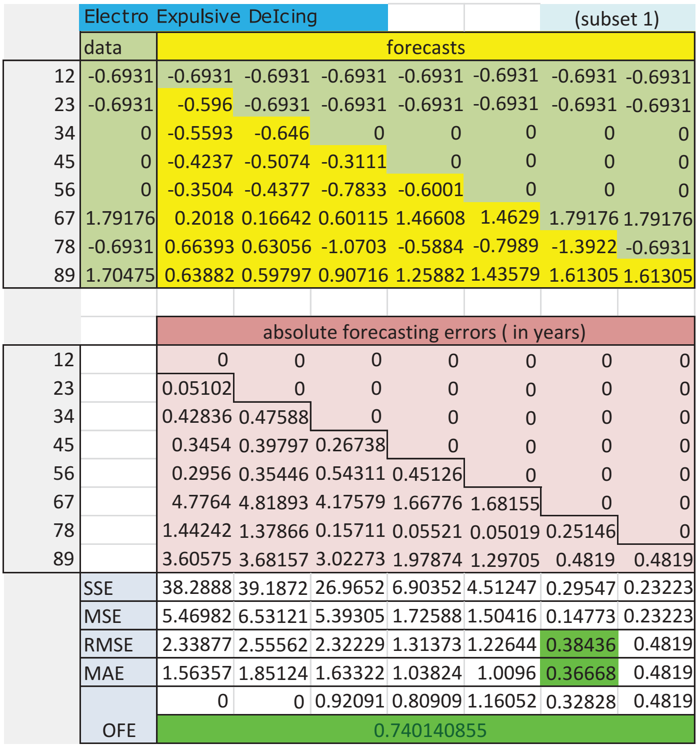
Other
- Closest Neighbor

Fixed estimates methods give the same forecast for all the technologies, they do not use past transitions to forecast future ones, and they constitute the reference measures. As a result, they are models that use the level-2 assumption, and they do not make use of the level-3 assumption. They will be used only as a reference to evaluate the performance of the other “smarter” level-3 techniques. Influence diagram methods (Shachter & Kenley, 1989) use a probabilistic approach by assuming a multivariate normal distribution of the variables and mapping the relations between them. Extrapolation techniques do not use training data; they are only based on the past transitions of the forecasted technology itself. Regression techniques are related to the influence diagram and extrapolation techniques. The forecasted transition is regressed against the known steps (using the training data), and the results are applied to the known past transitions in order to generate the forecast. Finally, the closest neighbor technique tries to forecast the transitions by imitating the variations of the technology in the training set that correlates the most with the technology being forecasted.

Performance of forecasting methods was evaluated using a multivariate comparison method. The two tables in Figure 4 show results for a single forecast. On top (in blue) is the name of the technology and the training subset, and the left column (in grey) shows the eight variables (TRL transitions). The upper table is the forecast table: The known values (i.e., past values) are in green, and the forecasted ones are in yellow. The different columns correspond to different known transitions: The first column makes forecasts for

TABLE 3 Correlation table of the NASA log-transition times (modified dataset)

Correlation Table	In(12) log data	In(23) log data	In(34) log data	In(45) log data	In(56) log data	In(67) log data	In(78) log data	In(89) log data
In(12)	1.000	0.660	0.752	0.312	0.149	−0.074	−0.135	−0.606
In(23)	0.660	1.000	0.905	0.673	0.385	0.043	−0.170	−0.350
In(34)	0.752	0.905	1.000	0.639	0.351	0.113	−0.256	0.265
In(45)	0.312	0.673	0.639	1.000	0.490	0.344	0.006	0.073
In(56)	0.149	0.385	0.351	0.490	1.000	0.325	0.331	0.307
In(67)	−0.074	0.043	0.113	0.344	0.325	1.000	−0.092	0.633
In(78)	−0.135	−0.170	−0.256	0.006	0.331	−0.092	1.000	0.180
In(89)	−0.606	−0.350	−0.265	0.073	0.307	0.633	0.180	1.000



	Electro Expulsive DeIcing						(subset 1)
	data	forecasts						
12	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931
23	-0.6931	-0.596	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931	-0.6931
34	0	-0.5593	-0.646	0	0	0	0	0
45	0	-0.4237	-0.5074	-0.3111	0	0	0	0
56	0	-0.3504	-0.4377	-0.7833	-0.6001	0	0	0
67	1.79176	0.2018	0.16642	0.60115	1.46608	1.4629	1.79176	1.79176
78	-0.6931	0.66393	0.63056	-1.0703	-0.5884	-0.7989	-1.3922	-0.6931
89	1.70475	0.63882	0.59797	0.90716	1.25882	1.43579	1.61305	1.61305

		absolute forecasting errors (in years)						
12		0	0	0	0	0	0	0
23		0.05102	0	0	0	0	0	0
34		0.42836	0.47588	0	0	0	0	0
45		0.3454	0.39797	0.26738	0	0	0	0
56		0.2956	0.35446	0.54311	0.45126	0	0	0
67		4.7764	4.81893	4.17579	1.66776	1.68155	0	0
78		1.44242	1.37866	0.15711	0.05521	0.05019	0.25146	0
89		3.60575	3.68157	3.02273	1.97874	1.29705	0.4819	0.4819
	SSE	38.2888	39.1872	26.9652	6.90352	4.51247	0.29547	0.23223
	MSE	5.46982	6.53121	5.39305	1.72588	1.50416	0.14773	0.23223
	RMSE	2.33877	2.55562	2.32229	1.31373	1.22644	0.38436	0.4819
	MAE	1.56357	1.85124	1.63322	1.03824	1.0096	0.36668	0.4819
		0	0	0.92091	0.80909	1.16052	0.32828	0.4819
	OFE	0.740140855						

FIGURE 4 Series of forecasts generated for one technology and different measures of error of those forecasts.

seven transitions knowing only the first one, while the last column makes forecasts for the last transition knowing all the previous ones.

The lower red table shows the absolute errors in years of every forecast, and it aggregates those errors in three different ways: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and OFE (Objective Function of Error). The OFE uses as inputs the "raw" error data that correspond to the 28 error values that appear in the red tables of the forecasts. The OFE aggregates this data based on "control parameters" such as the relative weight of forecast spans and whether or to consider or not a particular column. It also allows dropping a technology (mainly the outliers) out of the final total OFE.

Figure 5 compares the total OFE of all the methods using the median of the forecasting errors generated by Monte Carlo sampling with each curve corresponding to one forecasting technique. For a better visibility, we removed the worst techniques ("closest neighbor," "autoregression," and "regression").

The bounded autoregression (orange curve) method seems to be the best overall forecasting method. As for the rest of the methods, they perform better than fixed estimates in early stages and worse in the later stages. This conclusion on the other methods holds if we add the "extreme" outlier to the comparison (Figure 6).

Overall, the autoregression method can be considered the best method. Not only does it consistently outperform the other techniques for different OFE settings, its results also have a lower standard deviation. In this case, we went from an OFE of 13.2 for the best fixed estimate (the median) down to 11.2 for the autoregression (a 15% reduction in median

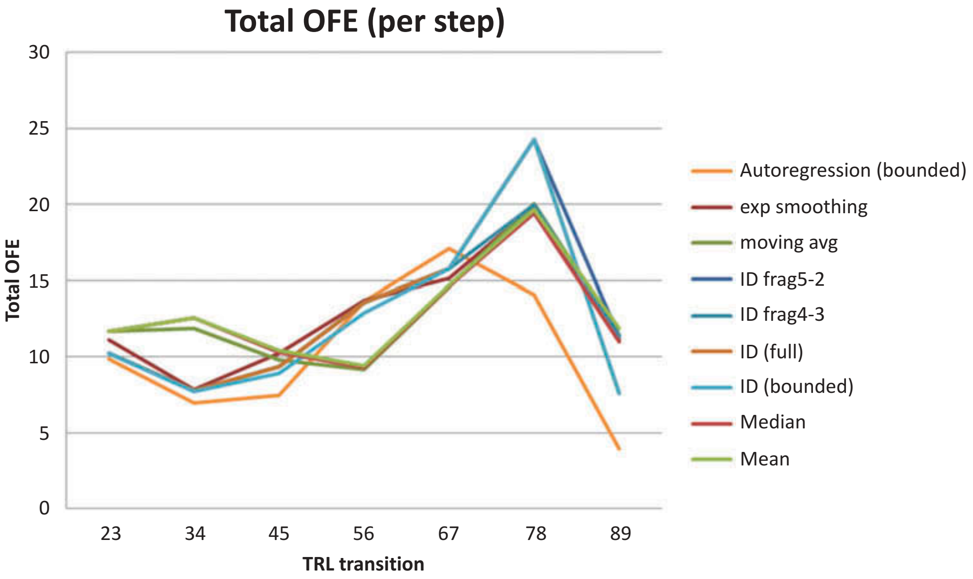


FIGURE 5 Total OFE for the nine best forecasting methods.

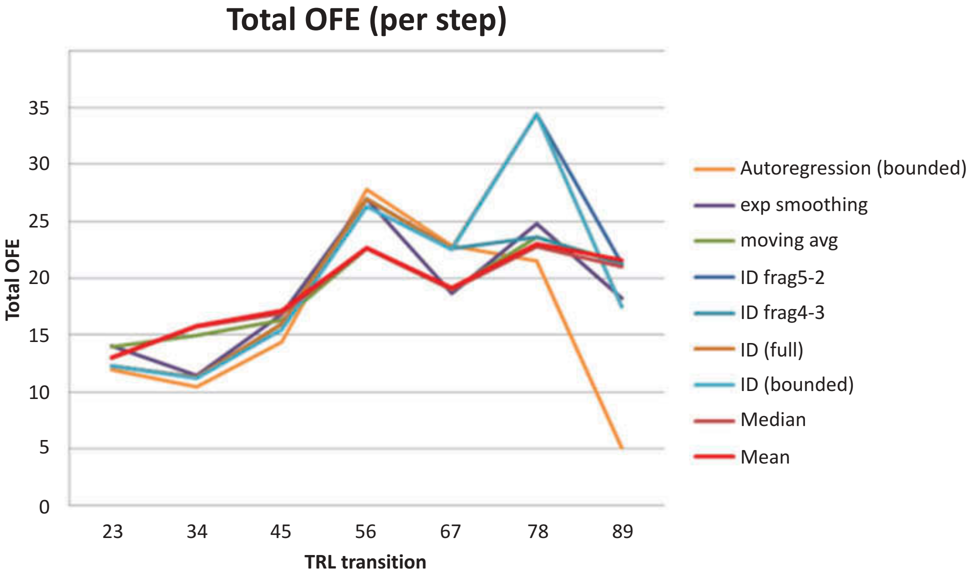


FIGURE 6 Total OFE for the nine best forecasting methods (with the extreme outlier included in the validation dataset).

forecast error). Most of the proposed techniques outperform the fixed estimates in the well-behaved TRL 1 to TRL 5 area. Then, all methods experience an increase in forecast error for the last transitions. While this poor performance can be attributed to the lack of data for those late transitions (only 11 data points available, of which only subsamples were taken for training), it is also a result of the high oscillation in transition times seen at those stages, or simply the higher variance of $X_{6–7}$, $X_{7–8}$, and $X_{8–9}$.

Although extra care was taken by resampling the dataset many times and looking only at median performances of the techniques, it is still possible that the methods have overlearned the dataset or got "lucky" for this specific sample. More data is needed to properly validate the approach to make sure that the improvement remains as the dataset gets larger. As a result, it might be better to use the robust median bootstrap technique of level 2 until level-3 forecasting techniques are validated for larger datasets.

Analysis of the Level-4 Assumption

The level-4 assumption states "TRL is a measure of maturity and risk." So far, we have not used TRL numbers as metrics; we have just used the fact that TRLs are ordered placeholders, defining distinguishable intervals, and that those intervals can be related. The fourth level in our framework gives meaning to the TRL numbers by stating that those numbers are a measure of maturity and risk.

This assumption can be decomposed into two parts. The first part is that TRL measures maturity. This part alone is of little use to managers since they are interested in the operational dimensions of maturity (cost risk, schedule risk, performance risk). Therefore, the second part of the assumption extends the first part to claim that the TRL numbers are also measures of those operational dimensions of maturity.

More specifically, risk analysts would like to use a direct relation between TRL and risk in the form of Risk = f (TRL). Many such models (for both cost and schedule) are available in the literature. Lee and Thomas (2001) regressed cost growth measures against a system TRL (average of the components weighted by their cost), while Dubos et al. (2008) regressed Relative Schedule Slippage against a truncated value of the cost-weighted TRL (WTRL). We found similar regressions against other maturity scales. Kenley and Creque (1999) performed a regression of Time-To-Maturity against the technical maturity (TM) scale, while Hoy and Hudak (1994) regressed cost uncertainty against a cost-weighted risk score.

In 2009, Boeing filed a patent on "systems, methods, and computer products for modeling a monetary measure for a good based upon technology maturity levels" (Matthews, Data, Feely, & Gauss, 2009). The patent has a broad scope, as it concerns any association of a TRL level with a component cost uncertainty, that is used to create component cost distributions, which are then added to generate the good's (i.e., the system's) cost distribution (Mathews, Datar, Feely, & Gauss, 2009).

While those models cannot be more valid than the assumptions they make, there is still room for improvement by correcting for the ordinality of the TRL scale and by being more careful with the relationship between TRL and risk. Doing so will allow us to answer the research question:

- Out of the available methodologies, what is the best one to use in performing regression of cost or schedule against TRL?

There are two weaknesses that show up often in the literature that are not weaknesses in the scale itself, but that emerge when the scale is used as a measure of technology risk.

The first weakness is that TRL is a bad measure for risk since it measures achieved developments only, without measuring the likelihood of a successful transition in the future. For instance, one technology can be at TRL 6, and still have a costly and long development ahead of it, while another technology can be at TRL 2, and have easy low-risk development ahead of it. Azizian et al. (2009), Cornford and Sarsfield (2004), Nolte (2008), Mahafza, Componation, and Tippett (2005), Sharif, Yu, and Stracener (2012), and Fernandez (2010) all pointed out this problem in the TRL scale. In response to this issue, Mankins (1998) and Bilbro (2007) respectively developed the AD^2 and the RD^3 measures. Those measures aim to quantify the difficulty and risk of what is left to be done, as opposed to what has already been done.

The second problem is that risk managers would like to use TRL as a measure of risk. However, TRL does not fully represent risk since it is only related to the probability of occurrence term, while it misses the other factor in risk, which is the consequence term. Hence, caution is required as to how TRL should be integrated in risk analysis. However, even if we make sure to use TRL to measure only the probability of occurrence, it is still a weak measure (especially when compared with other measures like RD^3 and AD^2, for the reasons explained above). This is why Conrow (2003) recommended never using TRL alone in risk assessments.

Many measures like RD^3 and AD^2 were specifically developed to compensate for TRL's failure to measure risk. They are intended to complement TRL in that they measure what TRL misses: future maturation risk. As a result, this assumption does not have theoretical support. Conrow (2009) noted that TRL is only weakly correlated with risk. This assumption is partially true, however, in the weak sense defined by the assumption for level 1. The risk decreases as TRL increases because the risk now is on a fewer number of remaining steps, and not because the TRL number itself is a measure of risk.

Although TRL does not measure technology risk, the literature indicates that TRL surprisingly explains a very large percentage of the variation in maturity variables with regression results with R^2 values at around 90% or above. Conrow (2011) contested those results by performing his own regression of Schedule Change on the data from Lee and Thomas (2001), and getting a R^2 of 0.26. Conrow (2009) also mentioned that R^2 should be adjusted for the degrees of freedom, and he obtained adjusted R^2 values of 0.52 and 0.02 when repeating the calculations of Dubos et al. (2008).

TRL is a partial measure of maturity risk. It is important to take the definition of the regression's dependent variable into consideration (e.g., total cost vs. absolute cost growth vs. relative cost growth vs. probability of cost growth). We did not have sufficient data to confirm a high correlation in all those cases, but evidence in the literature confirms a useful correlation between the level of TRL and the variation in maturity variables such as cost and schedule.

Conrow (2003, 2009, 2011) pointed out that risk is decomposed into a probability of occurrence term and a consequence of occurrence term. While TRL can be a partial measure of the probability of occurrence, it provides little information on the consequence of occurrence. If a technology has a low TRL when product development begins, we can say that there are high chances of cost overruns, but it is harder to estimate by how much the overruns will be. Therefore, if we look at cost risk for example as our relevant maturity parameter, then we could look at many cost-related variables. "Probability of cost growth" would be a variable that can be modeled well by using TRL (because it is a probability of occurrence term). On the other hand, "Absolute cost growth" is a variable that would be poorly modeled by TRL (because it is a consequence of occurrence term). "Relative cost growth" might have a chance of being well-modeled by TRL since it is a normalization of the consequence of occurrence.

Conrow (2003, 2009, 2011) also pointed out another issue: the ordinality of the TRL scale. The numbers in the TRL scale have no particular meaning; they are simple placeholders that just indicate the order, and could well be replaced by letters A, B, C, D, and so on with no loss of information. Mathematical operations (sums, averages, differences, multiplications) on TRL scores are not defined; they have no mathematical meaning, and can lead to wrong results. Such operations are nevertheless found in the literature.

Gorod, Sauser, and Boardman (2008) defined System Readiness Level (SRL), as a weighted sum of TRLs for system components that uses Integration Readiness Levels

(IRLs) to determine the weighting. SRL does not take into account the criticality of some components in the Work Breakdown Structure (WBS). Dubos et al. (2008) and Lee and Thomas (2001) both performed TRL averaging when creating the WTRL. First, averaging ordinal numbers does not have a mathematical meaning. Second, if the scale were cardinal, there would be no reason to weigh the average with component costs (cost does not necessarily represent the component's criticality in the system). Third, if the scale were cardinal, there would be no reason for averaging to be the appropriate operation to aggregate TRLs in a system. The minimum might be a better operation, and more generally, the WBS should be taken into account to detect parallelisms and critical paths.

Another example of performing mathematical operations is the Integrated Technology Index (ITI) proposed by Mankins (2002):

$$ITI = \frac{\sum_{subsystem\ technologies} \left(\Delta TRL \times R\&D^3 \times TNV\right)}{Total\ \#\ of\ subsystem\ technologies}. \tag{1}$$

The ITI disregards the ordinality of TRL twice: first when performing differences of TRLs and second when performing a weighted average of those numbers.

Furthermore, regressions on the TRL scale are over-constrained by the fact that TRL is treated as cardinal. A linear regression for example would be looking for a constant increment of the maturity variable between any two consecutive TRLs. This constraint has no reason to be true.

We are not implying that all results involving operations on TRL are necessarily wrong and misleading. The application might be small enough or defined on a small range of the TRL scale that the operation would not lead to major mistakes. However, caution should be used in defining the range of TRL values, and then they should be tested extensively to show that the range of results could carry useful information for the user.

As a solution to this problem, Conrow (2009) proposed a calibration of the TRL scale in an attempt to give a maturity meaning to the TRL numbers. Conrow asked experts to compare the maturity of the 36 pairs of TRL values and then aggregated and normalized the results using an Analytic Hierarchical Process (AHP). He then scaled the results so that the calibrated TRL 9 had a value of 9, and then fitted the results to a third degree polynomial. The result was:

$$TRL_{adjusted} = 0.346 + 0.012 * TRL^3. \tag{2}$$

While this calibration provides a technique to adjust TRL numbers to a metric scale, it still suffers a couple of drawbacks.

First, AHP is based on human assessments that are often imprecise and subjective, especially when using not-so-well defined words such as "maturity." Second, there is a conceptual problem in calibrating the scale with respect to "maturity." What does it mean for an expert to answer, "TRL 9 is 2.5 times more mature than TRL 6"? If we want to make an ordinal scale cardinal, then the calibration has to be done with respect to an underlying cardinal space. El-Khoury (2012) pointed out that the maturity concept is a multidimensional concept, and we cannot say that A is more mature than B by using only one number. For the comparison to make sense, it should pick one of the operational dimensions of maturity. For example, "What are the odds of meeting schedule for TRL 9 compared to TRL 6?" or "What is the ratio of cost overrun risk between TRL 8 and TRL 4?"

Nevertheless, the calibration still captures the general idea that the distance between maturity variables becomes larger at higher TRLs. One alternative approach to avoid the problem of ordinality in regressions is to treat the TRLs as categorical variables. Instead of embedding in advance a certain distance between the TRL levels in the regression, we can simply consider each TRL a different category and let the regression compare them by pairs.

In summary, while the problem of probability vs. consequence of occurrence and the problem of ordinality might be addressed by using specific methodologies, we still cannot solve the issue of TRL not being a measure of maturity, or it not being a measure of risk. This would necessitate a multiple regression against several maturity scales, which is beyond the scope of what we were able to analyze with the limited data sets. To answer our earlier question on the best methodology in performing regression against TRL, this methodology should:

- Avoid using any kind of averaging of TRLs. A WBS-based approach or SRLs can be used instead, if system TRLs need to be computed.
- Have a probability-of-occurrence-related dependent variable. The choice of variable is also important in that it has to be one that is well explained by the regression (in terms of adjusted R^2).
- Use a calibrated TRL scale (preferably calibrated to a relevant dimension of maturity).

Conclusions and Directions for Future Research

In this article, we proposed a four-level taxonomy of TRL-based models. When we were performing our research on TRL cost and schedule models, we discovered that different models departed from different assumptions and that the only way of consistently evaluating and comparing those models was by grouping them according to the increasing assumptions that they make. This allowed us to make recommendations and propose some alternative methodologies that make better use of the assumptions. Each level was analyzed theoretically (mainly by using properties of the TRL scale) and empirically (mainly by using the NASA dataset). The NASA dataset was also used in developing and evaluating the new methodologies at levels 2 and 3, as a well as in making recommendations at levels 1 and 4. Data scarcity was a major concern throughout our research, and we often had to use specific methodologies and propose special recommendations to deal with this fact.

The framework, although theoretical, allowed us to answer four practical research questions, one at each assumption level:

At level 1, the model strongly supported GAO maturation risk recommendations, both theoretically and empirically.

At level 2, the data confirmed that TRL transition times were statistically differentiated enough for the study of transition time distributions to make sense. Furthermore, the bootstrap method is a suitable technique to make estimations on those variables with small datasets.

At level 3, we found that there is significant positive correlation between early TRL transition times. Furthermore, the autoregression method improved forecast accuracy by 15% over level-2 fixed estimate models.

At level 4, the assumptions are only partially supported, although regressions on TRL still explain a significant percentage of variability in maturity variable risk. However, we propose some modifications to the methodologies found in the literature so that they avoid performing mathematical operations on TRLs, that they correct for TRL's ordinality, and that they limit the model to probability-of-occurrence terms only.

All those results allow us to say that TRL can indeed be used beyond simply the exit criteria of a Science and Technology program.

The presented models are in no case a silver bullet in solving the problems in acquisition. A lot of those problems can be mainly attributed to managerial issues, poor risk management practices, external changes in funding or requirements, or to wrong incentive structures (for example, the contracting structure). This main contribution of this research is to provide a framework to aid future researchers in understanding and developing more realistic cost and schedule models to improve the risk analysis part of risk management.

Many potential directions of research directly follow from the work presented in this article.

First, although special care was taken in choosing the methods and in making conclusions and recommendations, the empirical part of this thesis is still based on a very small dataset. The results need to be confirmed as soon as more data becomes available.

Second, although there are reasons to believe that the four assumptions can be extended to other types of technologies and other agencies, let us note that our empirical evidence only supports the assumptions for NASA technologies for now. The results in Crépin et al. (2012) are in support of a generalization of results to the Department of Energy's TRL, but they are also based on a small dataset and only on a few of the 8 TRL transitions. More statistical comparison work can be done as other agencies adopting TRL start releasing data.

Third, a very important extension to this research is to augment the models by looking at other complementary measures of maturity. We can try to capture a larger proportion of variability in maturity variables by performing multiple regression analysis.

Finally, there is still work to be done to improve level-4 regression models. Initially, we can find what versions of cost and schedule variables best correlate with TRL. In addition, we can develop and test calibrations of the TRL scale that are relevant on certain dimensions of maturity.

ORCID

C. Robert Kenley http://orcid.org/0000-0003-1350-5350

References

Albright, S. C., Winston, L. W., & Zappe, C. (2006). *Data analysis and decision making with Microsoft Excel* (3rd ed.). Mason, OH: South-Western Publication Co.

Azizian, N., Sarkani, S., & Mazzuchi, T. (2009). *A comprehensive review and analysis of maturity assessment approaches for improved decision support to achieve efficient defense acquisition*. Paper presented at the Proceedings of the World Congress on Engineering and Computer Science.

Bilbro, J. W. (2007). *Mitigating the adverse impact of technology maturity*. Paper presented at the NASA Proejct Management Challenge.

Conover, W. J. (1999). *Practical nonparametric statistics*. New York, NY: Wiley.

Conrow, E. H. (2003). *Effective risk management: Some keys to success*. Reston, VA: AIAA (American Institute of Aeronautics & Astronautics).

Conrow, E. H. (2009). Estimating technology readiness level coefficients. *AIAA Paper No. 6727*.

Conrow, E. H. (2011). Estimating technology readiness level coefficients. *Journal of Spacecraft and Rockets*, *48*(1), 146–152. doi:10.2514/1.46753

Cornford, S. L., & Sarsfield, L. (2004, 6–13 March 2004). *Quantitative methods for maturing and infusing advanced spacecraft technology*. Paper presented at the Aerospace Conference, 2004. Proceedings. 2004 IEEE.

Crépin, M., El-Khoury, B., & Kenley, C. R. (2012). *It's all rocket science: On the equivalence of development timelines for aerospace and nuclear technologies*. Paper presented at the INCOSE International Symposium, Rome, Italy.

De Angelis, D., & Young, G. A. (1992). Smoothing the bootstrap. *International Statistical Review/Revue Internationale de Statistique*, 45–56.

Department of Defense (DoD). (2009). Technology Readiness Assessment (TRA) Deskbook. Washington, DC: Director, Research Directorate (DRD). Office Of The Director, Defense Research And Engineering (DDR&E).

Dubos, G. F., & Saleh, J. H. (2011). Spacecraft technology portfolio: Probabilistic modeling and implications for responsiveness and schedule slippage. *Acta Astronautica*, *68*(7–8), 1126–1146. doi:10.1016/j.actaastro.2010.10.007

Dubos, G. F., Saleh, J. H., & Braun, R. (2008). Technology readiness level, schedule risk, and slippage in spacecraft design. *Journal of Spacecraft and Rockets*, *45*(4), 836–842. doi:10.2514/1.34947

El-Khoury, B. (2012). *Analytic framework for TRL-based cost and schedule models*. (Unpublished master's thesis.) Cambridge, MA: Massachusetts Institute of Technology. Retrieved from http://hdl.handle.net/1721.1/78484

Fernandez, J. A. (2010). *Contextual role of TRLs and MRLs in technology management*. Albuquerque, NM: Sandia National Laboratories.

Foden, J., & Berends, H. (2010). Technology management at Rolls-Royce. *Research-Technology Management*, *53*(2), 33–42.

GAO. (1999). *Best practices: Better management of technology development can improve weapon outcomes*. (NSIAD-99-162). Washington, DC: General Accounting Office.

Gorod, A., Sauser, B., & Boardman, J. (2008). System-of-systems engineering management: A review of modern history and a path forward. *Systems Journal, IEEE*, *2*(4), 484–499. doi:10.1109/JSYST.2008.2007163

Graettinger, C., Garcia, S., & Ferguson, J. (2003). *TRL corollaries for practice-based technologies*. Pittsburgh, PA: Carnegie Mellon University, Software Engineering Institute.

Hoy, K. L., & Hudak, D. G. (1994). Advances in quantifying schedule/technical risk. *Paper presented at the Analytical Sciences Corporation, 28th DoD Cost Analysis Symposium*, Xerox Document University, Leesburg, VA.

Johnson, N. L. (1949). Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, 36, 149–176.

Kenley, C. R., & Creque, T. R. (1999). Predicting technology operational availability using technical maturity assessment. *Systems Engineering*, *2*(4), 198–211. doi:10.1002/(SICI)1520-6858(1999)2:4<198::AID-SYS2>3.0.CO;2-7

Lee, T. S., & Thomas, L. D. (2001). Cost growth models for NASA's programs: A summary. *Computing Science and Statistics*, *33*, 431–440.

Lee, T. S., & Thomas, L. D. (2003). Cost growth models for NASA's programs. *Journal of Probability and Statistical Science*, *1*(2), 265–279.

Legresley, P. A., Bathke, T., Carrion, A., Cornejo, J. D., Owens, J., Vartanian, R.,. . . Kroo, I. M. (2000). 1998/1999 AIAA foundation graduate team aircraft design competition: Super stol carrier on-board delivery aircraft. *SAE transactions*, *109*(1), 1096–1111.

Mahafza, S., Componation, P., & Tippett, D. (2005). A performance-based technology assessment methodology to support DoD acquisition. *Defense Acquisition Review Journal*, *11*(13), 268–283.

Mankins, J. C. (1995). Technology readiness levels. Washington, DC: Advanced Projects Office, Office of Space Flight, NASA Headquarters.

Mankins, J. C. (1998). Research & Development degree of difficulty (R&D3). Washington, DC: Advanced Projects Office, Office of Space Flight, NASA Headquarters.

Mankins, J. C. (2002). Approaches to strategic research and technology (R&T) analysis and road mapping. *Acta Astronautica*, *51*(1–9), 3–21. doi:10.1016/s0094-5765(02)00083-8

Mathews, S. H., Datar, V. T., Feely, K., & Gauss, D. J. (2009). 8,204,775. U. S. P. Office.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference* (Vol. 95). Sage Publications, Incorporated.

Nolte, L. W. (2008). *Did I ever tell you about the whale? Or measuring technology maturity*. Information Age Publishing, Inc.

Olive, D. J. (2005). A simple confidence interval for the median. *Manuscript*. Retrieved from http://www.math.siu.edu/olive/ppmedci.pdf

Peisen, D. J., Schulz, C. L., Golaszewski, R. S., Ballard, B. D., & Smith, J. J. (1999). *Case studies: Time required to mature aeronautic technologies to operational readiness*. Draft Report. Arlington, VA: SAIC.

Sauser, B., Ramirez-Marquez, J. E., Magnaye, R., & Tan, W. (2008). A systems approach to expanding the technology readiness level within defense acquisition. *International Journal of Defense Acquisition Management*, *1*, 39–58.

Shachter, R. D., & Kenley, C. R. (1989). Gaussian influence diagrams. *Management Science*, *35*(5), 527–550. doi:10.2307/2632102

Sharif, A., Yu, J., & Stracener, J. (2012). *The U.S. Department of Defense technology transition: A critical assessment*. Paper presented at the 22nd Annual International Symposium of the International Council of Systems Engineering, Rome, Italy.

Smoker, R. E., & Smith, S. (2007). System cost growth associated with technology-readiness level. *Journal of Parametrics*, *26*(1), 8–38.

About the Authors

Bernard El-Khoury is a consultant with the Boston Consulting Group in Dubai, United Arab Emirates. He has a master's degree in Technology and Policy from the Massachusetts Institute of Technology, a master's degree in Industrial Engineering from Ecole Centrale Paris, and a bachelor's degree in engineering from Ecole Centrale Paris. His research interests are technology cost and schedule forecasting, and power systems modeling.

C. Robert Kenley is an Associate Professor of Engineering Practice in Purdue's School of Industrial Engineering in West Lafayette, Indiana. He has doctoral and master's degrees in Engineering-Economic System from Stanford University, a master's degree in statistics from Purdue University, and an bachelor's degree in management from the Massachusetts Institute of Technology. He has over thirty years' experience in industry, academia, and government as a practitioner, consultant, and researcher in systems engineering. He has published papers on systems requirements, technology readiness assessment and forecasting, Bayes nets, applied meteorology, and the impacts of nuclear power plants on employment.