An Integrated Quantitative Approach to Acceptance Testing and Related Decisions

John MacCarthy
University of Maryland
2175 A.V. Williams Building
College Park, MD 20742
301.405.4419
jmaccart@umd.edu

Abstract. This paper provides an integrated, analytical framework (and associated tools) for determining the scope of an acceptance test, designing acceptance tests, and for making acceptance decisions. A decision tree model is developed that may be used to determine the value of different acceptance test designs. Hypothesis testing models are developed that may be used to determine the performance threshold for testing, given a desired confidence of failing a bad system (test "Power"), and the design margin required to have a given confidence of passing a good system (test "Confidence"). These two models are then integrated into a single model that may be used to establish a test design (in terms of an acceptance threshold and the number of measurements) that achieves a desired test Power and Confidence. Finally a decision tree model is developed for the system acceptance decision based on the results of a given acceptance test.

Introduction and Approach

Verification testing and acceptance (validation) testing are two of the most important forms of testing on a program. The results of such testing are used to make decisions regarding continued funding of a program, proceeding to low-rate initial production (or full production), a need to correct deficiencies, or cancellation of a program. This paper provides an integrated, quantitative approach to planning acceptance (and verification) tests and for making an acceptance decision.

There are two principal questions associated with acceptance (and verification) testing. The first question is "How much testing is enough?" While Engel (2010) provides a rather complex framework for addressing this question, in this paper we will provide a simplified version of this approach (in the form of a decision tree) to obtain the "value" of a test (which may be used to answer the question of how much testing is enough).

The second question associated with acceptance testing is "What course of action should one take given the results of the test?" Guillaume-Joseph and MacCarthy (2017) have addressed the development of a decision tree model for the acceptance decision based on the results of an acceptance test. In this paper, we put this work within the context of the first question and the problem of validation (verification) test design.

The design of verification and acceptance tests is generally a rather complex activity. In recent years there has been particular interest in rigorous application of quantitative hypothesis testing and the Design of Experiments (DoE) to the design of tests. These techniques are addressed in texts on the experimental design and statistics (e.g., Montgomery 2012 and Navidi 2014).

One generally wants to design a test that is characterized by a high "Power" (probability of failing a bad system) and a high "Confidence" (probability of passing a good system). Two important test design parameters are what acceptance threshold (Xt) to set for each key performance parameter (KPP) and how many samples must be taken (Ns) to obtain a mean measured value for each KPP of interest (Xmm). The values that should be selected for Xt and Ns depend on the desired test Power and Confidence and on the required value for the KPP (Xr) and its expected value (Xe), based on analysis and/or previous testing. In this paper, we will run through an example of how this is done.

This paper draws heavily from exemplars developed as part of the University of Maryland's systems engineering master's program's core course in verification, validation, and testing.

Hypothesis Testing Background and Definitions

This section provides a set of definitions and variables that will be used throughout the paper. These definitions are based on a "supplier's perspective" to hypothesis testing (i.e., the "null hypothesis" is that the system is "good").

1. System is "Good" means that the system actually (over the long run, for all units) has a performance that meets or exceeds the required values established for all of its KPPs. The "actual" population mean will be denoted by Xai and the required values will be denoted by Xri.
2. System is "Bad" means that it fails to meet one or more required KPP values (over the long run).
3. A system "Passes" a test if the measured mean value obtained for each KPP in that test exceeds the pass criterion threshold value for that KPP. The test-measured mean value for each KPP will be denoted by Xmmi and the pass criteria (test threshold) values will be denoted by Xti.
4. A system "Fails" a test if one or more measured values fail to meet the pass criteria values.
5. The "Test (or Design) Margin" value of a KPP is the "expected" performance of the system (denoted by Xei). This is the value that one calculates based on the predicted performance of the design and/or the system's performance in previous tests. In case where a high value of a KPP is desired, one generally wants this value to be greater than the test's pass criteria values in order have a greater than 50 percent confidence that the system will pass the test. Alternatively, if low values of the KPP are desired, one wants Xei < Xti.
6. P(G) is the "prior" probability that the system is "good." It may be calculated from a system simulation based on the system's design and/or from the results of previous system tests.
7. P(P|G) is the conditional probability that the system will pass the test, given it is good.
8. P(F|G) is the conditional probability that the system will fail the test, given it is good, i.e., the likelihood of a Type I error.
9. P(F|B) is the conditional probability that the system will fail the test, given it is bad.
10. P(P|B) is the conditional probability that the system will pass the test, given it is bad, i.e., the likelihood of a Type II error.

11. The "Confidence" of a test (C) is the probability that the system will pass the test, given the system is good, i.e., C = P(P|G). Note that this assumes the supplier's perspective on what constitutes the "null hypothesis" (i.e., the system is "Good").
12. The "Power" of a test (Pr) is the probability that the system will fail the test, given the system is bad, i.e., C = P(F|B). Again, this is from the supplier's perspective.

Figure 1 provides the "Hypothesis Testing Matrix" that identifies the probabilities associated with each possible state of reality and each possible test outcome.

Test Result	Reality	
	System is Good	System is Bad
Passes Test	P(P|G) = C = 1-α	P(P|B) = β (Type II Error)
Fails Test	P(F|G) = α (Type I Error)	P(F|B) = Pw = 1-β

Figure 1. Hypothesis Testing Matrix

Calculating the Value of an Acceptance Test

This section of the paper uses a decision tree approach to answer the question "How much testing should be done?"

In order to develop our simple model for the value of an acceptance test, we need to make the following simplifying assumptions (and associated definitions).

1. There are three possible courses of action (COAs) available to the decision maker:
   - COA 1 = Do not test the system.
   - COA 2 = Provide a "quick" test of the system.
   - COA 3 = Provide an "extensive" test of the system.
2. The expected total costs of each of these options are Cnt, Ct1, and Ct2, respectively.
3. There is only one type of (serious) fault.
4. From previous tests and/or design analysis we can calculate the prior probability that the system has the serious fault (is bad), i.e., P(B).
5. We can calculate the probability that test design i will find a fault, given it exists (P(Fi|B)) from the test design (see the Determining the Performance Margin section).
6. We can calculate the probability that test design i will pass a system, given no fault exists (P(Pi|G)) from the test design (see the Determining the Nm and Xt section).
7. If the fault is found, it can and will be fixed.
8. The cost of fixing a fault depends on when it is found. Specifically, let Cfft be the cost of fixing a fault found during testing, while Cfff is the cost of fixing a fault once the system is fielded (generally much greater).
9. The "value of a test"i may be calculated using Vti = Cnt – Cti.
10. For the illustrative reference case, we will assume: P(B) = 0.3, P(Fo|B) = 0, P(F1|B) = 0.2, P(P1,G) = 0.9, P(F2|B) = 0.6, P(P2,G) = 0.9, Cnt = $0, Ct1 = $1 M, Ct2 = $4 M, Cfft = $5 M, and Cfff = $100 M.

Based on these assumptions, we may develop a decision tree model for determining which acceptance testing strategy has the least total cost. This decision tree is provided in Figure 2.

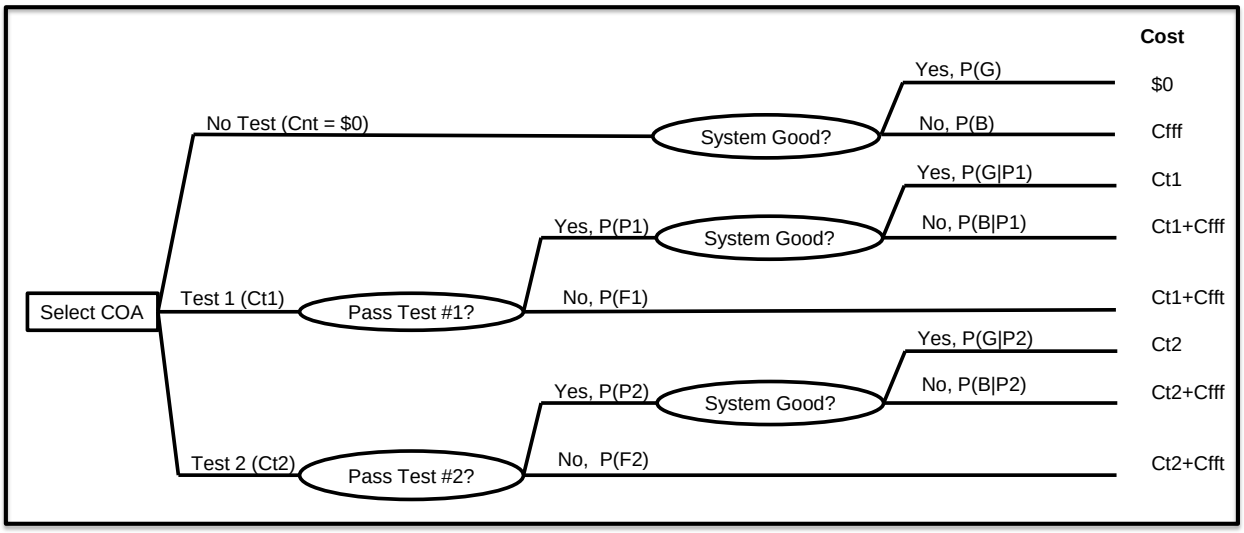

Figure 2. Decision Tree for Three Testing Options

In order to solve this decision tree, we need to calculate the indicated probabilities using the following information and relationships:

- We are given P(B), P(F|B), P(P|G), and the Costs.
- We know P(G) = 1-P(B) and P(P|B) = 1- P(F|B).
- We also know P(Pi) = P(Pi|G) P(G) + P(Pi|B) P(B) and P(Fi) = 1-P(Pi).
- And we can use Bayes' Theorem to get P(B|Pi) = P(Pi|B) * P(B) / P(Pi).
- And we know P(G|Pi) = 1-P(B|Pi).

From these relationships and the decision tree, we can calculate the Expectation Value for the Total Cost associated with each decision option as follows:

- **Cnt = EV(COA1)** = P(B) * Cfff = 0.3 * $100 M **= $30 M**
- **Ct1 = EV(COA2)** = Ct1 + P(P1) * P(B|P1) * Cfff + P(F1) * Cfft = 1 + 0.24*100+0.13*5 **= $25.7 M**
- where:
  - P(P1) = P(P1|G) P(G) + P(P1|B) P(B) = 0.9*0.7+0.8*0.3 = 0.87
  - P(B|P1) = P(P1|B) * P(B) / P(P1) = 0.8 * 0.3 / 0.87 = 0.276
  - Ct2 = EV(COA3) = Ct2 + P(P2) *P(B|P2) * Cfff + P(F2)*Cfft = 4 + 0.12*100+0.25*5 = $17.3 M
  - P(P2) = P(P2|G) P(G) + P(P2|B) P(B) = 0.9*0.7+0.4*0.3 = 0.75
  - P(B|P2) = P(P2|B) * P(B) / P(P2) = 0.4 * 0.3 / 0.75 = 0.16

These results lead us to the conclusion that even though Test 2 is four times as expensive as Test 1, it has a significantly lower total expectation cost than either not testing or executing Test 1. Finally, it follows from the model that the value of Test 2 is

V2 = Ct2-Cnt = $12.7 M.

In the next two sections, we will turn our attention to how to develop a test design that will provide desired values for P(Fi|B) and P(Pi|G).

Determining the Acceptance Test Performance Threshold to Achieve a Desired Test Power

This section of the paper addresses two related test design questions: (1) "What should be selected as the pass criteria threshold (Xt) for a given KPP in order to be confident that the test will fail a bad system (one for which Xa<Xr)?" and (2) "How many measurements (Nm) need to be taken?"

In order to address this question, we provide the following simplifying assumptions and definitions.

1. X is the metric of interest, and high values of X are desired.
2. Let SDm be the standard deviation in measurements of X, and we have an estimate of this from previous tests (or analysis).
3. Let SEm be the standard error in the mean obtained for a set of Nm measurements. Note that:

$$SEm = \frac{SDm}{\sqrt{Nm}}. \quad (1)$$

4. Let σa be the standard deviation associated with "true" (population) mean.
5. Assume σa ~ SDm.
6. Assume measured values of X (Xm) will reflect a random sample of the true (population) value of Xa.
7. Assume that the actual system performance is Xa = Xr-δ (where δ~0).

Recall from the Calculating the Value of an Acceptance Test section that the Power of a test (Pw) is the probability of failing a bad system (P(F|B). The Power of a test is a function of Xr, Xt, the number of measurements taken (Nm), and the standard deviation associated with Xa (σa ~ SDm). The problem may be expressed graphically as shown in Figure 3.

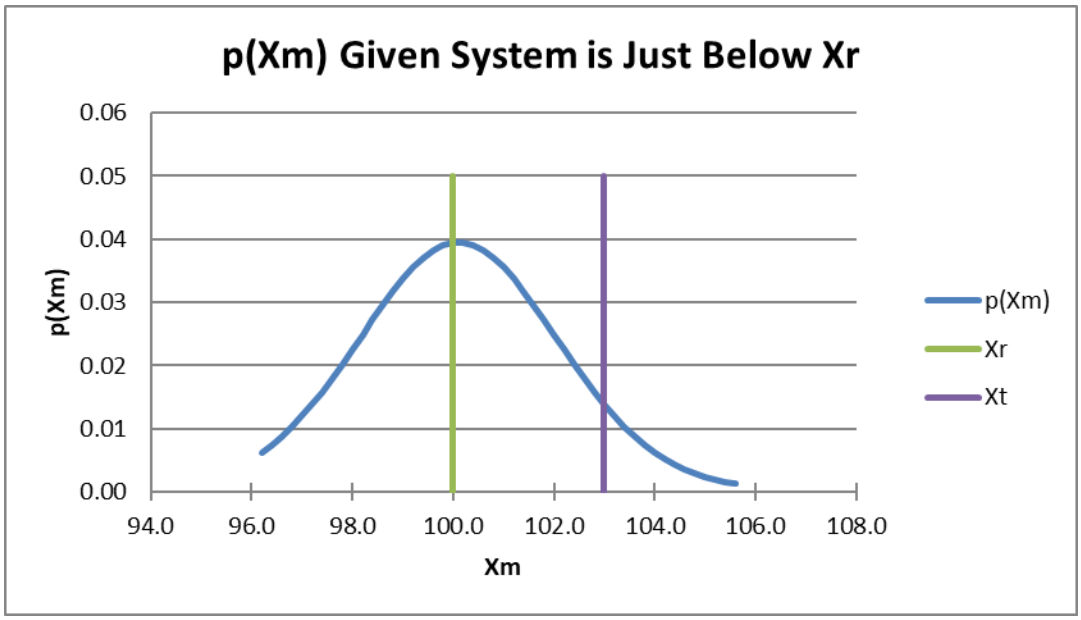

Figure 3. Probability of Obtaining a Measured Mean Xm, given Xa = 100 Units

Figure 3 shows the probability density function associated with the t distribution for the measured mean (Xmm), given that the actual value of the metric is just a hair lower than the required value (i.e., Xa ~< Xr = 100 units), the standard error is SEm = 2 units (from a standard deviation SDm = 10 units and Nm = 25), and the test pass criteria threshold value is Xt = 103 units.

This example allows us to calculate the Power (Pw) associated with this test design (Nm = 25 and Xt = 103 units). To do so, we calculate the "t statistic" of these values of Xt, Xr, and SEm using:

$$t(\beta) = \frac{(Xt - Xr)}{SEm} = 1.5. \quad (2)$$

We can solve for β using the Excel @T.DIST.RT(t, Nm-1) function (or a Student's t distribution table). Doing this, we obtain β = 0.073, which indicates that the Power of this test is Pw = 0.937 (i.e., we are ~94 percent confident that we will fail a bad system).

With regard to test design, if we knew we only required a Power of 0.9, we could reduce Xt, or reduce Ns, or reduce both. To determine the degree to which we could reduce these test design parameters, we rewrite equation (2) as:

$$Xt = Xr + t(1 - Pw) * SEm. \quad (3)$$

Using the Excel @T.INV(Pw,Nm-1), we get t = 1.32. From this, we can determine that if we keep Nm = 25 (and SEm = 2), we could reduce Xt to Xt = 100+2.6 = 102.6 units. Alternatively, if we were to reduce Nm to 16, SEm would increase to SEm = 2.5 units and Xt would become Xt = 100+3.3 = 103.3 units.

Note that in the case where Nm > 30, the t-distribution may be approximated by the normal distribution, and one uses the Z transformation rather than the t transformation, where z replaces t in

equation (3), and the Nm-independent Excel @NORM.S.DIST(z,1) function is used to find Pw or the @NORM.S.INV(Pw) function is used to find z (and thus Xt).

In the next section, we will see how to determine what performance is required to achieve a desired Confidence for passing a good system.

Determining the Performance (Design) Margin Required to Pass the Acceptance Test

This section of the paper addresses the question of "What performance margin is required in order to be confident that the test will be passed, if the system is good?"

In this analysis, we make use of simplifying assumptions 1-3 from the previous section and the assumption that the actual system performance is equal to the expected system performance, i.e., Xa = Xe.

Recall from the Calculating the Value of an Acceptance Test section that the Confidence of a test (C) is the probability of passing a good system (P(P|G). The Confidence of a test is a function of Xd, Xt, the number of measurements taken (Nm), and the standard deviation associated with Xm (SDm).

The problem is illustrated in Figure 4, where we have plotted the probability density function associated with the t distribution for the measured mean (Xmm), given that the actual value of the metric is the expected value (i.e., Xa = Xe = 107 units). The standard error is SEm = 2 units (from a standard deviation SDm = 10 units and Nm = 25), and the test pass criteria threshold value is Xt = 103 units.

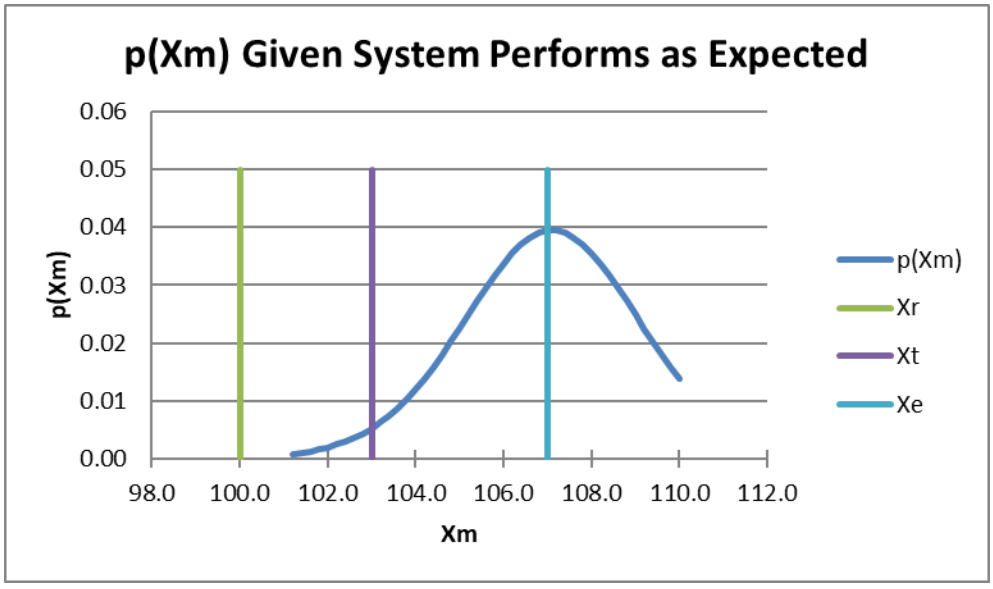

Figure 4. Probability of Obtaining a Measured Mean Xm, given Xa = Xe = 107 Units

This example allows us to calculate the Confidence (C) associated with this test design (Nm = 25 and Xt = 103 units) using the same approach described in the previous section, but with the following modifications. To do so, we calculate the "t statistic" of these value of Xt, Xr, and SEm using:

$$\mathrm{t}(\alpha) = \frac{(\mathrm{Xe}-\mathrm{Xt})}{SEm} = 2. \tag{4}$$

As in the previous section, we solve for α using the Excel @T.DIST.RT(t, Nm-1,1) function (or a Student's t distribution table). Doing this, we obtain β = 0.028, which indicates that the Confidence of this test is C = 0.972 (i.e., we are ~97 percent confident that we will pass a good system).

With regard to test design, if we knew we only required a Confidence of 0.9, we could reduce the needed design margin (Xe), or reduce Ns, or reduce both. To determine the degree to which we could reduce these test design parameters, we rewrite equation (4) as:

$$Xe = Xt + \mathrm{t}(1-C) * SEm. \tag{5}$$

Using the Excel @T.INV(C,Nm-1) function, we get t = 1.32.

From this, we can determine that if we keep Nm = 25 (and SEm = 2), we could reduce Xe to Xe = 103+2.6 = 105.6 units. Alternatively, if we were to reduce Nm to 16, SEm would increase to SEm = 2.5 units, and Xe would become Xt = 103+3.3 = 106.3 units.

Again, if Nm > 30, the t-distribution may be approximated by the normal distribution and one uses the Nm-independent Excel @NORM.S.DIST(z,1) function to find C or the @NORM.S.INV(C) function to find Xe.

In the next section, we will see how to combine the results of this and the previous section to develop a test design based on given values Pw, C, Xr, Xe, and SDm.

Determining the Nm and Xt from Pw, C, Xr, Xe, and SDm

In general, the problem of test design involves determining how many measurements need to be taken (Nm) and what value to set for the pass criteria threshold Xt in order to achieve desired values for Pw and C. In general, one will know the value for Xr and have rough estimates for Xe and SDm (from previous testing and/or analysis).

The solution to this problem involves solving equations (3) and (5) simultaneously (we have two equations and two unknowns, Nm and Xt). Substituting the expression for Xt in equation (3) and for SEm in equation (1) into equation (5) we obtain:

$$Xe = Xr + [\mathrm{t}(1-Pw) + \mathrm{t}(1-C)] * \frac{SDm}{\sqrt{Nm}}. \tag{6}$$

Rearranging terms, we get:

$$\frac{Nm}{[\mathrm{t}(1-Pw)+\mathrm{t}(1-C)]^2} = \left[\frac{SDm}{Xe-Xr}\right]^2. \tag{7}$$

While an exact solution to equation (7) requires the use of a numerical method (t is a non-linear function of Nm), one may obtain a closed form approximate solution in the case where Nm > 30. In this case, the t-distribution may be replaced by the normal distribution and we obtain:

$$Nm = [\mathrm{z}(1 - Pw) + \mathrm{z}(1 - C)]^2 * [\frac{SDm}{Xe - Xr}]^2. \quad (8)$$

As an example of the use of equations (8) and (3), consider the case where Pw = 0.9, C = 0.8, Xr = 100 units, Xe = 108 units, and SDm = 25 units. Using equation (8) we obtain

Nm = (1.28+0.84)^2 * (25/8)^2 = 44 measurements.

Plugging this into equation (3) (and using z rather than t), we obtain:

$$Xt = 100 + z(0.1) * \frac{25}{\sqrt{44}} = 104.8 \text{ units}.$$

In short, we have shown that, given desired values of Pw, C, and Xr and expected values Xe and SDm, we may determine the key test design parameters of Xt and Xm.

The Acceptance Decision

In this section, we establish a decision tree model that can provide a quantitative framework for the acceptance decision.

Generally, there are three acceptance decision courses of action available to a decision maker: (1) accept the system, (2) fix and retest the system, or (3) reject the system. Each of these decision options can have very different financial consequences depending on whether the system was adequately able to perform its mission ("good") or not ("bad"). The results of an acceptance test are usually a major consideration in this decision.

In developing our decision model we make the following simplifying assumptions and definitions.

1. There are three possible courses of action (COAs) available to the decision maker:
    - COA 1 = Accept the system.
    - COA 2 = Fix and retest the system.
    - COA 3 = Reject the system.
2. If the system passes the first acceptance test, it will be accepted.
3. Ci is the expectation value for the total cost associated with COAi.
4. The results of the acceptance test provide an estimate of the probability that the system is good P(G) = 1 - γ, where $\mathrm{t}(\gamma) = \frac{(\mathrm{Xmm} - \mathrm{Xr})}{SEm}$.
5. There is a lost opportunity cost (Clo) associated with failing to deploy (i.e., rejecting) the system.
6. There is a significant cost (Cfff) associated with having to fix the system if it needs to be fixed following deployment (see the Determining the Acceptance Test Performance Threshold section).

7. There is a much lower cost associated with fixing (and retesting) the system prior to deployment (Cft).
8. The probability of fixing the system post-deployment is 1.0.
9. Let Pf be the probability that the failed system can be fixed (prior to retesting).
10. Let P(P2) be the probability of passing the second acceptance test (and P(F2) is the probability of failing it).
11. Let P(P2|Fx) be the probability that the system will pass the second acceptance test, given it is fixed.
12. Let P(Fx|P2) be the probability that the system is fixed, given it passed the second acceptance test.
13. The fixed system will be accepted if it passes the second acceptance test.
14. For the illustrative reference case, we will **assume that the system failed the first acceptance test** and P(G) = 0.6 (based on the measured Xmm), Pf = 0.9 (from some analysis), (P2|Fx) = 0.9 and P(F2|~Fx) = 0.8 (same as for the first acceptance test), Cfff = $100 M, Cft = $5 M, and Clo = $200 M.

Given these assumptions, we construct the decision tree provided in Figure 5.

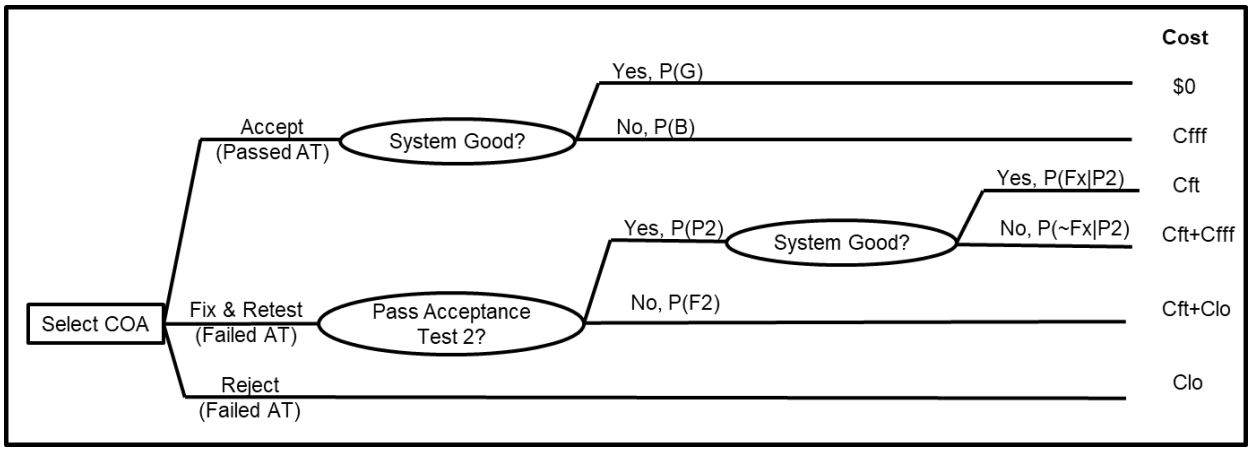

Figure 5. Decision Tree for the Acceptance Decision

In order to solve this decision tree, we need to calculate the indicated probabilities using the following information and relationships:

- We are given P(G), Pf, P(P2|Fx), P(P2|~Fx), and the costs.
- We know P(B) = 1-P(G).
- We know P(P2) = P(P2|Fx) * Pf + P(P2|~Fx) * (1-Pf) and P(F2) = 1 – P(P2).
- We can use Bayes' Theorem to get P(Fx|P2) = P(P2|Fx) * Pf/P(P2), and P(2) is found using the previous bullet.
- And we know P(~Fx|P2) = 1 - P(Fx|P2).

From these relationships and the decision tree, we can calculate the Expectation Value f**o**r the Total Cost associated with each decision option as follows:

- **C2 = EV(COA2)** = Cft + P(P2) * P(~Fx|P2) * Cfff + P(F2) * Clo

= 5 + 0.83*0.024*100 + 0.07*200 **= $21.0 M**
using P(P2) = P(P2|Fx) * Pf + P(P2|~Fx) * (1-Pf) = 0.9 * 0.9 + 0.2*0.1 = 0.83
and P(F2) = 1-P(P2) = 0.07
and P(Fx|P2) = P(P2|Fx) * Pf/P(P2) = 0.9*0.9/0.83 = 0.976
and P(~Fx|P2) = 1-P(Fx|P2) = 0.024

- **C3 = EV(COA3) = $200 M**

Based on this analysis of the reference case, one would advise the decision maker to adopt COA 2 (fix and retest the system), since it has a lower expected total cost than either accepting the system or cancelling the program. One should note the relatively low probability that the system was good (0.6, obtained from the first acceptance test) and that the high cost of having to fix a fielded system made acceptance too high a risk.

Conclusions

This paper provided an integrated, analytical framework (and associated tools) for determining the scope of acceptance and verification testing, designing acceptance and verification tests, and for making acceptance decisions.

It provided a decision tree model that may be used to determine the value of different acceptance (or verification) test designs. It developed a set of hypothesis testing-based models that can be used to determine (1) the performance threshold for acceptance (or verification testing) testing, given a desired confidence of failing a bad system (the "Power" associated with the test design); (2) the design margin required to have a given confidence of passing a good system (the "Confidence" associated with the test design); and (3) an acceptance test design (in term of an acceptance threshold and the number of measurements that must be taken) that achieves a desired test Confidence and Power. Finally, it provided a decision tree model for the system acceptance decision based on the results of a given acceptance test. Simple examples were also provided that illustrated the application of the framework and the use of the models.

It is hoped that the acceptance testing and decision framework and tools developed in the paper will prove useful to industry and government decision makers and test organizations and to those teaching courses that address system verification, validation, and testing.

References

Engel, A 2010, *Verification, Validation, and Testing of Engineered Systems*, Wiley, Hoboken, NJ (US).

Guillaume-Joseph, G and MacCarthy, J 2017, 'Performing Programmatic Trade-Off Analyses', in G Parnell (ed), *Trade-off Analytics: Creating and Exploring the System Tradespace,* Wiley, Hoboken, NJ (US).

Montgomery, D 2012, *Design and Analysis of Experiments*, 8th edn., Wiley, Hoboken, NJ (US).

Navidi, W 2014, *Statistics for Engineers and Scientists*, 4th edn.,McGraw-Hill Education, New York, NY (US).

Biography

Dr. John MacCarthy is currently the Director of the University of Maryland's Systems Engineering Education Program. Prior to this he held a variety of systems engineering leadership roles with TRW, Northrup-Grumman, and the Institute for Defense Analyses developing and evaluating large, complex systems and systems of systems. He also served for eight years as an Adjunct Professor of Systems Engineering at the University of Maryland, Baltimore County and five years as an Assistant Professor of Physics at Muhlenberg College. Dr. MacCarthy holds a B.A. in Physics from Carleton College, a Ph.D. in Physics from the University of Notre Dame, and an M.S. in Systems Engineering from George Mason University.