Result Evaluation

Document metadata: LO3 3.4 EVI 1

Contents:

Evaluation Criteria

- **Accuracy:** specifies how accurate the test is, i.e. how often it the correct ratio of true positives and true negatives.
- **Precision:** specifies how often the test identifies true positives out of all positive results.
- **Repeatability:** how easy it is to repeat the test (cost, effort, time, … more factors play into this).
- **Cost:** how expensive is it to execute this test (e.g. database read/writes).
- **Time:** how long does it take to execute this test?
- **Ease of use:** refers to ease of understanding, extending the test, eventually adopting it to system changes if necessary, and executing it.

In this testing environment, we do not have any training data to predict actual accuracy / precision / ... values. For those, we use 'n/d' (not determinable). For others, a rating on a scale is more valuable (e.g. repeatability). The scale we use ranges from 1 (very poor) to 5 (excellent). For instance, a 5 in "cost" is excellent, i.e. very low execution cost.

Evaluation of Results

Test	Accuracy	Precision	Repeatability	Cost	Time	Ease of use	Usability of results	Test-Specific Adequacy Criterion / Satisfy?
T1	n/d	n/d	4	5	0.9s	2.5	4 (more confidence in safe system)	
T4	n/d	n/d	2	3	230s	4	2 (no concurrent simulation so not much confidence in result)	
T6	n/d	2 (low sample size)	1	3	5min	2	5 (very useful usability feedback which led to direct system changes)	
T10	100%	100%	5	5	6s	5	3 (as stated before, compilation ability is only somewhat useful factor)	
T12	n/d	n/d	4	5	6s	2.5	4 (confidence in API protection)	
T15	n/d	n/d	5	5	<1s	5	5 (important confidence in unit logic)	
T18	n/d	n/d	5	5	<6s	5	5	statement coverage / yes
Avg.	**n/a**	**n/a**	**3.71**	**4.43**	**~79s**	**3.71**	**4**	