Data collection

Alex Calis

Allen Roginsky

April 27, 2021

Outline

- 90B requirements on Data Collection
- Summary of Data Collection requirements
- 6 Things to consider
- Questions?

90B requirements on Data Collection

- 3.1.1 – *Data Collection*
- 3.2.1 (items 5 & 6) – *Requirements on the Entropy Source*
- 3.2.4 – *Requirements on Data Collection*

90B Section 3.1.1 – *Data Collection*

1. A sequential dataset of at least 1,000,000 sample values obtained directly from the noise source (i.e., raw data) **shall** be collected for validation. If the generation of 1,000,000 consecutive samples is not possible, the concatenation of several smaller sets of consecutive samples (generated using the same noise source) is allowed. Smaller sets **shall** contain at least 1,000 samples. The concatenated dataset **shall** contain at least 1,000,000 samples.
2. If the entropy source includes a conditioning component that is not listed in Section 3.1.5.1.1, a conditioned sequential dataset of at least 1,000,000 consecutive conditioning component outputs **shall** be collected for validation. The output of the conditioning component **shall** be concatenated in the order in which it was generated and treated as a binary string for testing purposes. Note that the data collected from the noise source for validation may be used as input to the conditioning component for the collection of conditioned output values.
3. For the restart tests (see Section 3.1.4), the entropy source must be restarted 1,000 times; for each restart, 1,000 consecutive samples **shall** be collected directly from the noise source. The restart data **shall** be extracted whenever the noise source is ready and able to provide data that can be used for producing entropy source output. This data is stored in a 1,000×1,000 restart matrix M, where M[i][j] represents the j^{th} sample from the i^{th} restart.

90B Section 3.2.1 – *Requirements on Entropy Source*

5. When a conditioning component is not used, the output from the entropy source is the output of the noise source, and no additional interface is required. In this case, the noise source output is available during both validation testing and normal operation. When a conditioning component is included in the entropy source, the output from the entropy source is the output of the conditioning component, and an additional interface is required to access the noise-source output. *In this case, the noise-source output **shall** be accessible via the interface during validation testing, but the interface may be disabled otherwise. The designer **shall** fully document the method used to get access to the raw noise source samples. If the noise-source interface is not disabled during normal operation, any noise-source output using this interface **shall not** be provided to the conditioning component for processing and eventual output as normal entropy-source output*.
6. The entropy source may restrict access to raw noise source samples to special circumstances that are not available to users in the field, and the documentation **shall** explain why this restriction is not expected to substantially alter the behavior of the entropy source as tested during validation.

90B Section 3.2.4 – *Requirements on Data Collection*

1. The data collection for entropy estimation **shall** be performed in one of the three ways described below:
   - By the submitter with a witness from the testing lab, or
   - By the testing lab itself, or
   - Prepared by the submitter in advance of testing, along with the following documentation: a specification of the data generation process, and a signed document that attests that the specification was followed.

2. Data collected from the noise source for validation testing **shall** be raw output values.

3. The data collection process **shall not** require a detailed knowledge of the noise source or intrusive actions that may alter the behavior of the noise source (e.g., drilling into the device).

90B Section 3.2.4 – *Requirements on Data Collection*

4. Data **shall** be collected from the noise source and any conditioning component that is not listed in Section 3.1.5.1.1 (if used) under normal operating conditions.

5. Data **shall** be collected from the entropy source under validation. Any relevant version of the hardware or software updates **shall** be associated with the data.

6. Documentation of the data collection method **shall** be provided so that a lab or submitter can perform (or replicate) the collection process at a later time, if necessary.

7. Documentation explaining why the data collection method does not interfere with the noise source **shall** be provided.

Summary of Data Collection requirements (1/2)

Data **shall** be collected from:

1. The raw noise source
    1. At least 1 million samples.
    2. For restart tests (at least 1000 times with 1000 samples each time).
2. Any non-vetted conditioning component.
    1. At least 1 million samples each.
3. A well-defined interface or special needs software/hardware interface.
4. The actual entropy source under validation.

Data is collected by:

1. The lab directly or by the submitter (with lab observation or without observation with sufficient documentation).

Summary of Data Collection requirements (2/2)

Documentation **shall** include:

1. The method used to get access to the raw noise source samples.

2. The data collection method to demonstrate how to replicate the collection process.

3. Information explaining why the data collection method does not interfere with the noise source.

4. Why the restriction on accessing raw data (if applicable) is not expected to substantially alter the behavior of the entropy source.

Things to consider (1/6)

As mentioned already, access to raw noise is required, but what are some possible access methods? An entropy source may:

1. Have a well-defined interface to access raw bits.
   - This is the easiest method.
   - You just draw bits from the defined interface and justify why the data collection method does not interfere with the noise source.
2. Require a special needs access to raw bits.
   - This is more complicated.
   - Likely managed by the vendor/device maker.
   - The submitter would need to describe exactly how those bits were retrieved from the device and justify why this process doesn't alter the behavior of the noise source and gets the same noise source values that are used internally by the device.

Things to consider (2/6)

Plan ahead!

- You can’t test your design or verify anything about your model if you can’t get access to those raw bits.
- For example, it’s common to have a noise source that samples bits, and then keeps the XOR sum of many output bits. You can do this in hardware very efficiently. If you design in a mechanism to turn the XORing off, you have an easy way to get access to the raw noise source bits; otherwise, you don’t, and you’re going to have a pretty big challenge to get the raw bits from the source.

Things to consider (3/6)

Hardware devices usually run faster than the software collecting outputs for the test.

- For this reason, it's likely that the raw bits collected for validation will be a somewhat sparse subset of the bits generated internally. That is, you may very well collect 32 bits of raw outputs, and during the time they are output to whatever system is collecting them, may lose another 128 bits of raw outputs.
- The submitter needs to document this if it's happening and give a justification for why this won't invalidate the entropy estimators.

Things to consider (4/6)

What is the “raw source” when some designs may have multiple components that can be sampled?

- For example, a source that uses three ROs sampled on a stable clock and XORs the result together to generate an output bit, may be designed to allow sampling a single bit at a time or all three together.
- What the raw source is here depends on the model used for the source—if the designer has provided a model of the XORed result (e.g. part of digitization), then the raw bit may be considered for testing; if modeling the three internal bits, then those should be sampled and the XOR treated as a separate internal conditioning step.
- Need to explain where the entropy is coming from, how it’s measured, and justify what value should be considered a raw noise source output.

Things to consider (5/6)

For the restart tests, it is acceptable for the device to have some delay before it starts producing raw bits.

- This must be the same delay as is used in the fielded device.
- For example, if the sample device undergoing the restart tests delays half a second after resetting before it generates bits, the same must be true of the fielded device. Otherwise, the restart tests would not be meaningful.

Things to consider (6/6)

There's some overlap with the data collection and health tests.

- In general, health tests need to be done on the raw output bits, or on internal values that never produce an output. So, the designer already needs to have worked out how to get the raw bits for the continuous health tests—getting access to them for validation should hopefully be made easier by this.

Questions?

CMVP: cmvp@cmvp@nist.gov & CMVP@cyber.gc.ca

Alex Calis: alexander.calis@nist.gov

Allen Roginsky: allen.roginsky@nist.gov.