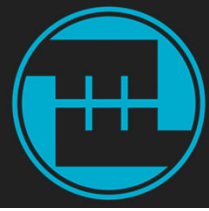
[Draft IG D.K] Comments

Joshua E. Hill, PhD

CMUF Entropy WG
20240319
Presentation version Draft 9

Some comments on the Draft IG D.K [Draft IG D.K, Resolution 22]

"If a sample size larger than 8 bits is used as input to the Adaptive Proportion Test (APT), the submitter **shall** include a justification that the assurances provided by the APT are still met by the health test. Namely the assurance that a drop in entropy from H to H/2 is detected with at least a 50% chance when observing the number of samples defined by the window size, W, chosen. Otherwise, an entropy source may map the larger sample size down to a sample size that is 8 bits or shorter specifically for running the APT. The method for mapping the data (e.g., truncation) and why this method does not hide health test failures shall be provided in the EAR."

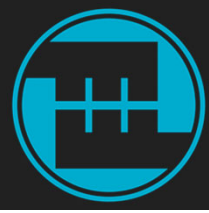


Some comments on the Draft IG D.K [Draft IG D.K, Resolution 22]

Part 1: “If a sample size larger than 8 bits is used as input to the Adaptive Proportion Test (APT), the submitter **shall** include a justification that the assurances provided by the APT are still met by the health test.”

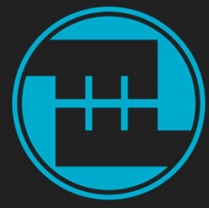
Part 2: “Namely the assurance that a drop in entropy from H to H/2 is detected with at least a 50% chance when observing the number of samples defined by the window size, W, chosen.”

Part 3: “Otherwise, an entropy source may map the larger sample size down to a sample size that is 8 bits or shorter specifically for running the APT. The method for mapping the data (e.g., truncation) and why this method does not hide health test failures shall be provided in the EAR.”



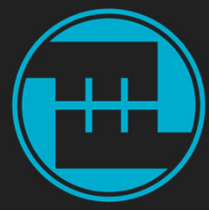
Part 1: Comments

- This is essentially a requirement that, under certain circumstances (when the raw data is suitably wide), the vendor/laboratory be able to demonstrate that the APT is doing something useful.
- This is an unconventional use of the term “sample size”, which normally refers to the number of observations in a data set. I’ll try to use “raw data width” in these comments.
- This appears to be conceptually related to SP 800-90B Section 4.3 Requirement 8 / IG D.K Resolution 14 (Shall IDs #77, #118, and #119, all of which are presently marked optional).



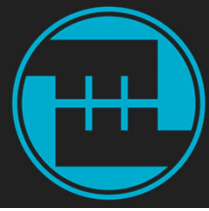
Part 1: Important APT Characteristics

- The APT is categorical, so how data is encoded (e.g., the raw data width) isn't relevant.
- The distribution parameter that establishes the cutoff in SP 800-90B Section 4.4.2 is the probability of the most likely symbol, $p_{\max}$ ($p_{\max} = 2^{-H}$ for an IID noise source). Here, we discuss this probability in terms of the "apparent entropy":
  - $H_{\text{apparent}} = -\log_2 p_{\max}$
  - $H \leq H_{\text{apparent}}$ (in all sources because of the MCV estimator).



Part 1: Important APT Characteristics

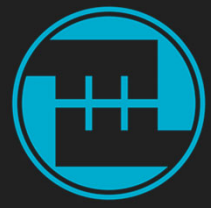
- For low apparent entropies, the number of symbols that can be output doesn't have much impact, as the most probable symbol is very likely to occur.
- For high apparent entropies, the residual probability must be associated with some symbols, so the observed APT false positive / statistical power are distribution-dependent.



Applicability: The Issue

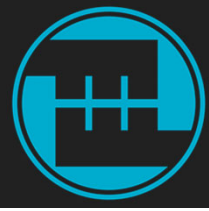
It isn't clear why this requirement should be restricted to wide raw data.

- Wide raw data doesn't have **any** impact on the APT, but APT cutoffs created based on high apparent entropy (i.e., low probability for the most likely symbol) may result in the APT having reduced statistical power.
  - An apparent entropy > 8 bits can only happen with raw data wider than 8 bits, but a noise source that produces wide raw symbols need not have a high apparent entropy.
  - A high apparent entropy requires small APT cutoffs, which eventually leads to a reduction in statistical power, and may not be compatible with the full false positive recommended range.
  - The "APT cutoff is too small" issue.



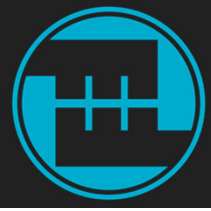
Applicability: The Issue

- Conversely, a very low apparent entropy (i.e., a very high probability for the most likely symbol) also causes problems.
  - A low apparent entropy requires a cutoff close to the window size, which eventually leads to a reduction in statistical power.
    - The SP 800-90B Section 4.4.2 procedure can yield a cutoff larger than the window size, which renders the APT useless (this is addressed by IG D.K Resolution 16).
  - An “APT cutoff is too large” issue.



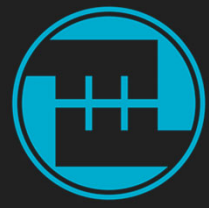
Applicability: The Issue

- The APT (using the Section 4.4.2 cutoff selection procedure) has a much smaller than targeted false-positive rate (and corresponding reduced statistical power) any time the assessed entropy isn't consistent with the apparent entropy.
  - Another "APT cutoff is too large" issue.
  - This is a significant issue for almost all non-IID sources.



Applicability: Possible Solution

- Remove the “sample size” (raw data width) applicability entirely.
  - The width of the raw data isn’t directly relevant.
- Rephrase the requirement so that it addresses both the “APT cutoff is too small” and “APT cutoff is too large” issues.

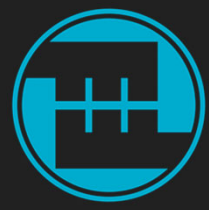


Some comments on the Draft IG D.K [Draft IG D.K, Resolution 22]

Part 1: “If a sample size larger than 8 bits is used as input to the Adaptive Proportion Test (APT), the submitter shall include a justification that the assurances provided by the APT are still met by the health test.”

Part 2: “Namely the assurance that a drop in entropy from H to H/2 is detected with at least a 50% chance when observing the number of samples defined by the window size, W, chosen.”

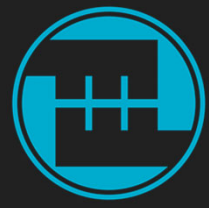
Part 3: “Otherwise, an entropy source may map the larger sample size down to a sample size that is 8 bits or shorter specifically for running the APT. The method for mapping the data (e.g., truncation) and why this method does not hide health test failures shall be provided in the EAR.”



Part 2: Comments

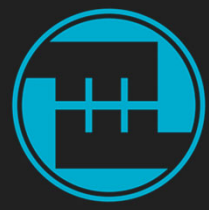
I see two problems with this part.

- The failure mode isn't completely specified.
- The required statistical power is much too large.



The Failure Mode: The Issue

- The statement "a drop in entropy from H to H/2" doesn't really specify what failure mode should be examined.
  - There are infinite ways this could happen.
  - What failure mode is relevant here, exactly?
    - How does the vendor/laboratory know if NIST will accept the failure mode they use for this analysis?
- If the vendor/laboratory isn't required to analyze the noise source's known or suspected failure modes (SP 800-90B Section 4.3 Requirement 8 / Shall ID #77 is optional) then how can they make this argument?

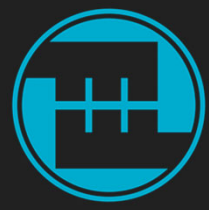


The Failure Mode: Possible Solution #1

- As this is essentially a modification of the SP 800-90B Section 4.5 Criterion (b) requirement, we could import that requirement's specific failure mode:

  "If the noise source's behavior changes so that the probability of observing a specific sample value increases to at least $P^* = 2^{-\frac{H}{2}}$…"

  - Advantage: This is very specific and easy to model.
  - Advantage: This is the failure mode that the APT was designed to detect.
  - Disadvantage: This isn't really a way that most non-IID noise sources degrade (barring a total failure, which the RCT would detect).

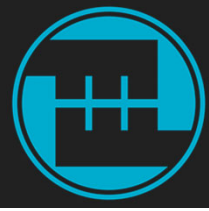


The Failure Mode: Possible Solution #2

- Leave the failure mode *more* unspecified:

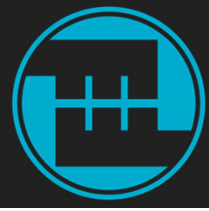
  "… justify why the APT can be expected to detect a failure mode…"

  - Advantage: This maximizes the vendor/laboratory flexibility to make a technical argument.
  - Disadvantage: It isn't obvious to the program participants what the requirement practically means.
  - Disadvantage: Absent some inter-laboratory coordination venue, it is likely that the various laboratory interpretations will become contradictory.



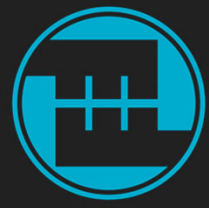
The Failure Mode: Possible Solution #3

- We could require (at least part of) SP 800-90B Section 4.3 Requirement 8 / IG D.K Resolution 14.
  - For this approach, the vendor/laboratory would need to describe at least one known or suspected failure mode that allows for a half-entropy failure mode.
- This would allow the vendor/laboratory to estimate the APT's statistical power when the entropy has fallen to a half-entropy condition due to the (selected) known or suspected failure mode.



The Failure Mode: Possible Solution #3

- Advantage: This known or suspected failure mode would be entropy-source specific.
- Disadvantage: Determining appropriate failure modes requires a detailed understanding of the design.
- Disadvantage: In many entropy sources, the APT isn't expected to detect the known or suspected failure modes that result in entropy reduction.
  - This suggests that the full approach specified in SP 800-90B Section 4.3 Requirement 8 / IG D.K Resolution 14 (Shall IDs #77, #118, and #119) would be a more comprehensive way to proceed.



Statistical Power: The Issue

IG D.K draft:

> "…is detected with at least a 50% chance when observing the number of samples defined by the window size, W, chosen."

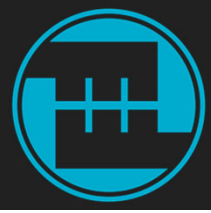
- This is a per-window statistical power requirement.

vs.

SP 800-90B Section 4.5, Criterion (b):

> "…the test shall detect this change with a probability of at least 50 % when examining 50 000 consecutive samples from this degraded source."

- This is a requirement for the statistical power across many windows.

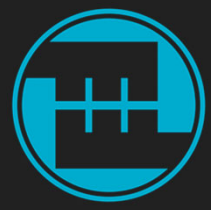


Statistical Power: The Issue

The per-window statistical power, β, required by SP 800-90B Section 4.5 Criterion (b) is a function of the window size.

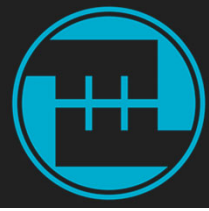
$$\beta \geq 1 - \left(\frac{1}{2}\right)^{1/\left\lfloor \frac{50{,}000}{W} \right\rfloor}$$

Window Size (W)	Minimum Per-Window Statistical Power (β)
512	0.7%
1024	1.4%



Statistical Power: The Issue

- The draft IG D.K per-window statistical power is 34-71 times the statistical power required by Criterion (b).
- The APT bounds produced by SP 800-90B Section 4.4.2 using the recommended false positive range $\alpha \in [2^{-40}, 2^{-20}]$ don't always have the required per-window statistical power while in a half-entropy failure mode.
  - Requiring this statistical power can foreclose using false positive rates in the recommended range.
- It doesn't make any sense that the statistical power required for the approved APT be so dramatically different than the requirements imposed on a developer-specified health test.

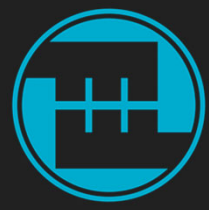


Statistical Power: Possible Solution #1

- As this is essentially a modification of the SP 800-90B Section 4.5 Criterion (b) requirement, we could import that requirement's specific failure mode:

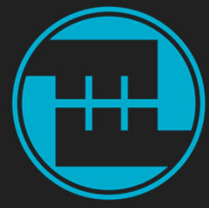
  "If the noise source's behavior changes so that the probability of observing a specific sample value increases to at least $P^* = 2^{-\frac{H}{2}}$…"

  - Advantage: This is very specific and easy to model.
  - Advantage: This is the failure mode that the APT was designed to detect.
  - Disadvantage: This isn't really a way that most non-IID noise sources degrade (barring a total failure, which the RCT would detect).



Statistical Power: Possible Solution #2

- Leave the required statistical power unspecified.
  - Advantage: This maximizes the vendor/laboratory flexibility to make a technical argument.
  - Disadvantage: It isn't obvious to the program participants what the requirement practically means.
  - Disadvantage: Absent some inter-laboratory coordination venue, it is likely that the various laboratory interpretations will become contradictory.

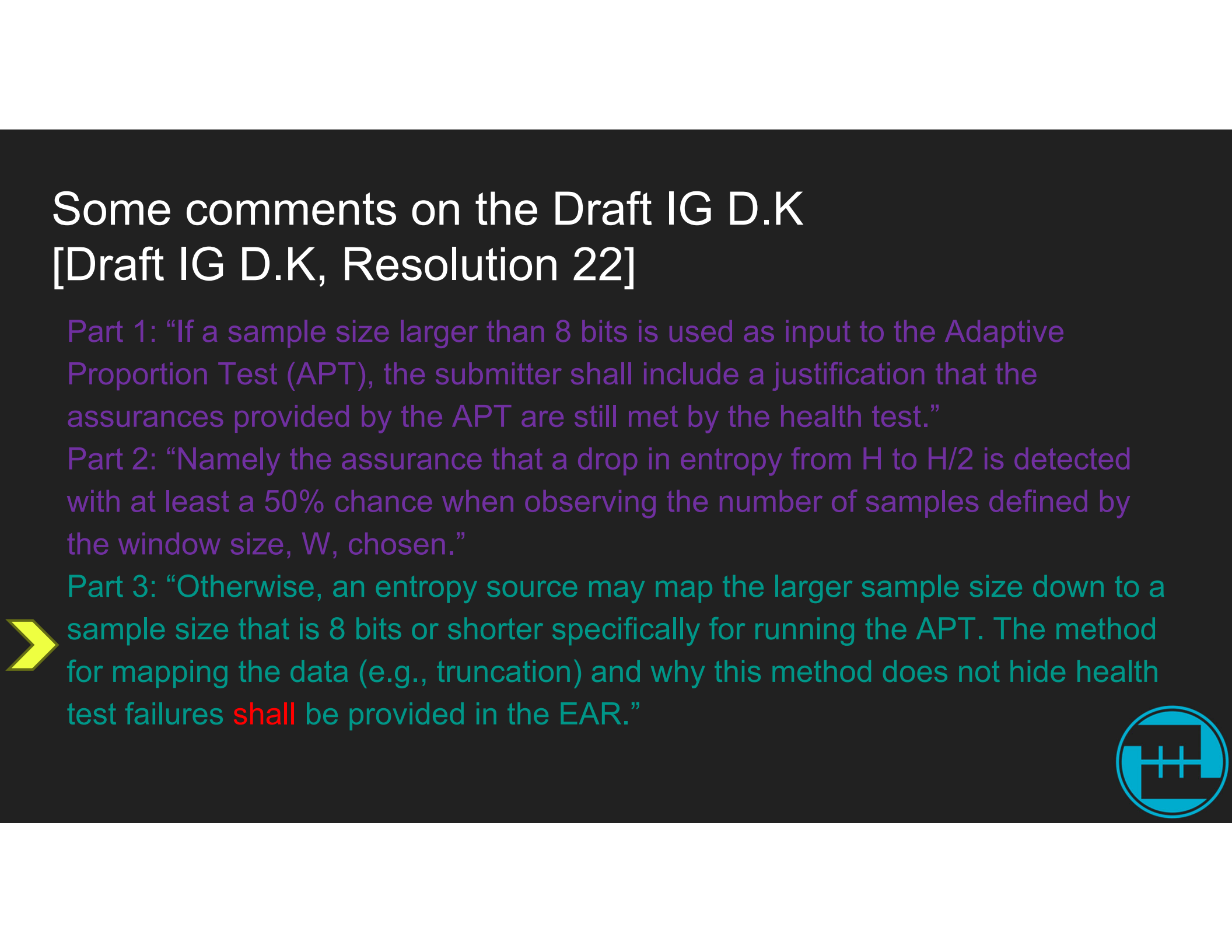


Some comments on the Draft IG D.K [Draft IG D.K, Resolution 22]

Part 1: “If a sample size larger than 8 bits is used as input to the Adaptive Proportion Test (APT), the submitter shall include a justification that the assurances provided by the APT are still met by the health test.”

Part 2: “Namely the assurance that a drop in entropy from H to H/2 is detected with at least a 50% chance when observing the number of samples defined by the window size, W, chosen.”

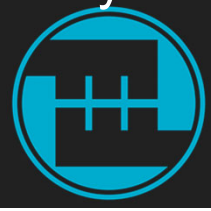
Part 3: “Otherwise, an entropy source may map the larger sample size down to a sample size that is 8 bits or shorter specifically for running the APT. The method for mapping the data (e.g., truncation) and why this method does not hide health test failures shall be provided in the EAR.”



Mapping and the APT: The Issue

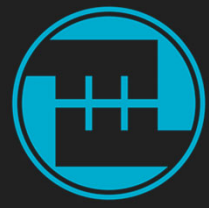
"… map the larger sample size down to a sample size that is 8 bits or shorter specifically for running the APT."

- Performing an APT on mapped data tends to obscure shifts in the probability distribution of the unmapped data.
- If we use the same APT cutoff in the mapped-data APT, then the mapped-data APT necessarily has the same or better statistical power, but also likely has a larger false positive rate.
- We are instructed by SP 800-90B Section 4.4.2 to select an APT cutoff based on the targeted false positive rate, so presumably after mapping we follow the same procedure. Such a (larger) mapped-data APT cutoff will result in a similar false positive rate for the mapped-data APT but may also result in a dramatically worse statistical power.



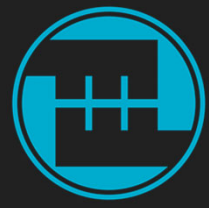
A Relevant Aside on JEnt

- NIST is particularly interested in how the Jitter Entropy Library (JEnt) APT performs, as internally JEnt uses a large (64-bit) raw data width.
- In practice, the delta format used may require a large integer to represent it (particularly with the idiomatic JEnt 3.0.1 and earlier delta format), but the number of symbols present is smaller than the raw data width suggests.
- The way that the symbols are encoded isn't relevant to the APT behavior (the APT is categorical).



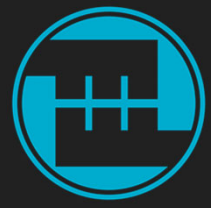
A Relevant Aside on JEnt

- JEnt 3.1.0 and later sets the APT cutoff based on the `osr`.
  - $H_{\text{submitter}} = \frac{1}{\texttt{osr}}$.
  - The target false positive rate is $\alpha = 2^{-30}$.
  - In earlier versions of JEnt, this cutoff is effectively fixed at 326 (erroneously off-by-one from the SP 800-90B Section 4.4.2 value for $H = 1.0$ for $\alpha = 2^{-30}$) or the APT is absent.



A Relevant Aside on JEnt

- For modern versions, this behavior is consistent with the SP 800-90B Section 4.4.2 procedure, but it is important to note that $\mathrm{H} \leq H_{\mathrm{submitter}} \ll H_{\mathrm{apparent}}$.
  - The observed false positive rate is dramatically less than the targeted false positive rate of $\alpha = 2^{-30}$ (this is an “APT cutoff is too large” issue).
  - The impact on the statistical power is dependent on the identified failure mode and the raw data distribution.

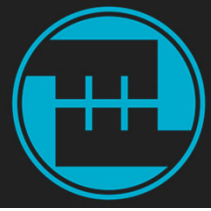


JEnt: An Example

Using JEnt 3.4.1 on a modern-ish Intel platform (Cascade Lake).

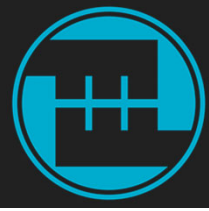
- A data set of 1 billion raw samples has about 25k distinct symbols (fewer than 2^{15}).
- By default, the `osr` is set to 3 on this platform, so here $\mathrm{H} = H_{\text{submitter}} = \frac{1}{3}$.

Entropy Estimate Type	Min Entropy	APT Cutoff for $\alpha = 2^{-30}$
Assessed Entropy	0.333333	459
Apparent Entropy	7.68651	18
8-Bit Mapped Apparent Entropy	7.41904	19
4-Bit Mapped Apparent Entropy	3.99920	71



A Relevant Aside on JEnt

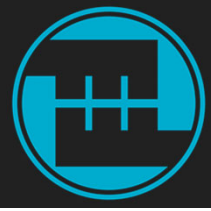
- Given a raw data sample from the noise source in a failure mode, we can estimate the maximum cutoff that attains a targeted failure rate.
- The distribution underlying the particular failure mode being examined is relevant.
- For this example, we simulated these failure modes by repeatedly replacing randomly selected data samples in the unmapped data set with the MLS until the desired rate was attained.
- Any translation then occurs on this full-width failure mode raw data.



JEnt: An Example

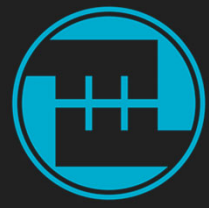
- What cutoff is required to detect various half-entropy modes at the targeted statistical powers? (Cutoff bounds in red are not met by the selected cutoffs.)

	Failure Mode Apparent Entropy	Maximum APT Cutoff for $\beta = 0.7\%$	Maximum APT Cutoff for $\beta = 50\%$
Unmapped Half Assessed	0.166667	473	455
Mapped to 8 bits Half Assessed	0.166492	473	455
Mapped to 4 bits Half Assessed	0.156459	473	455
Unmapped Half Apparent	3.84326	44	2
Mapped to 8 bits Half Apparent	3.82424	44	2
Mapped to 4 bits Half Apparent	3.01632	44	2



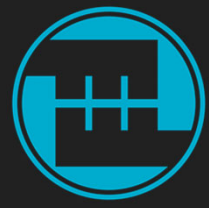
JEnt APT: You Get What's on the Label

- In most cases, the APT did well when detecting the failure mode it was configured to detect.
- Conversely, if you configure the APT based on an entropy assessment that is less than half the apparent entropy, then this APT isn't likely to detect a half-apparent-entropy failure mode.



JEnt: A Conclusion

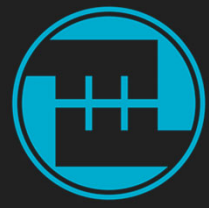
- The 50% targeted statistical power wasn't attained in any of the failure modes.
- For these distributions, mapping from 22k symbols to 256 symbols didn't have a substantial impact on the statistical power of the test for a fixed APT cutoff, though it did require a slightly larger cutoff to attain the same targeted false positive rate (and using this larger cutoff would reduce the statistical power).
- Mapping to less than the apparent entropy caused problems in the half-apparent-entropy failure modes.
- There was no observed benefit to mapping to a narrower raw data width.



Some Bound Computations

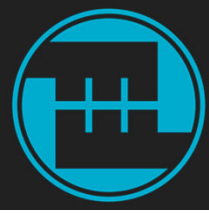
The following example computations are based on the SP 800-90B Section 4.4.2 analysis approach and formulas.

- There are relevant corrections ([HJ 2019, Comment 10b]), but they do not change the essential results.
- The SP 800-90B Section 4.4.2 analysis approach:
  - Makes an underlying IID assumption.
  - Bounds the false positive rate (α) and statistical power (β) using the assumption that the APT reference symbol (A) is the most likely symbol.
- The actual false positive rate and statistical power associated with a particular APT cutoff selection are distribution-dependent and can be estimated via large-scale simulation.



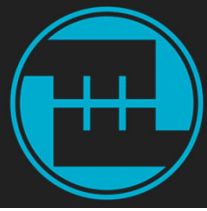
Mapping and the APT: Bounds

- We'll examine an IID source whose output follows the near-uniform distribution from [HD 2012], and which produces all possible 16-bit output symbols with an assessed 10 bits of min entropy.
- SP 800-90B Section 4.4.2 directs us to choose the window size $W = 512$ and an APT cutoff between 8 ($\alpha = 2^{-20}$) and 12 ($\alpha = 2^{-40}$).
- In the Criterion (b) half-entropy failure mode, our statistical power is over 99% for $\alpha = 2^{-20}$ and over 92% for $\alpha = 2^{-40}$ using the unmapped data.



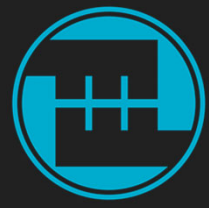
Mapping and the APT: An Example

- If we map the raw data by truncating it to 4 bits, then the entropy for the mapped data would be about 3.98 bits. For a 512-symbol window ($W = 512$) the APT cutoffs are then 63 ($\alpha = 2^{-20}$) to 79 ($\alpha = 2^{-40}$).
- The Criterion (b) half-entropy failure mode produces a mapped entropy of about 3.45 bits.
  - This half-entropy failure mode doesn't halve the mapped entropy!
  - The dramatically more probable pre-mapped symbol is now less obvious after mapping.
- Our statistical power is now about 1.5% for the $\alpha = 2^{-20}$ cutoff and essentially 0 for the $\alpha = 2^{-40}$ cutoff.

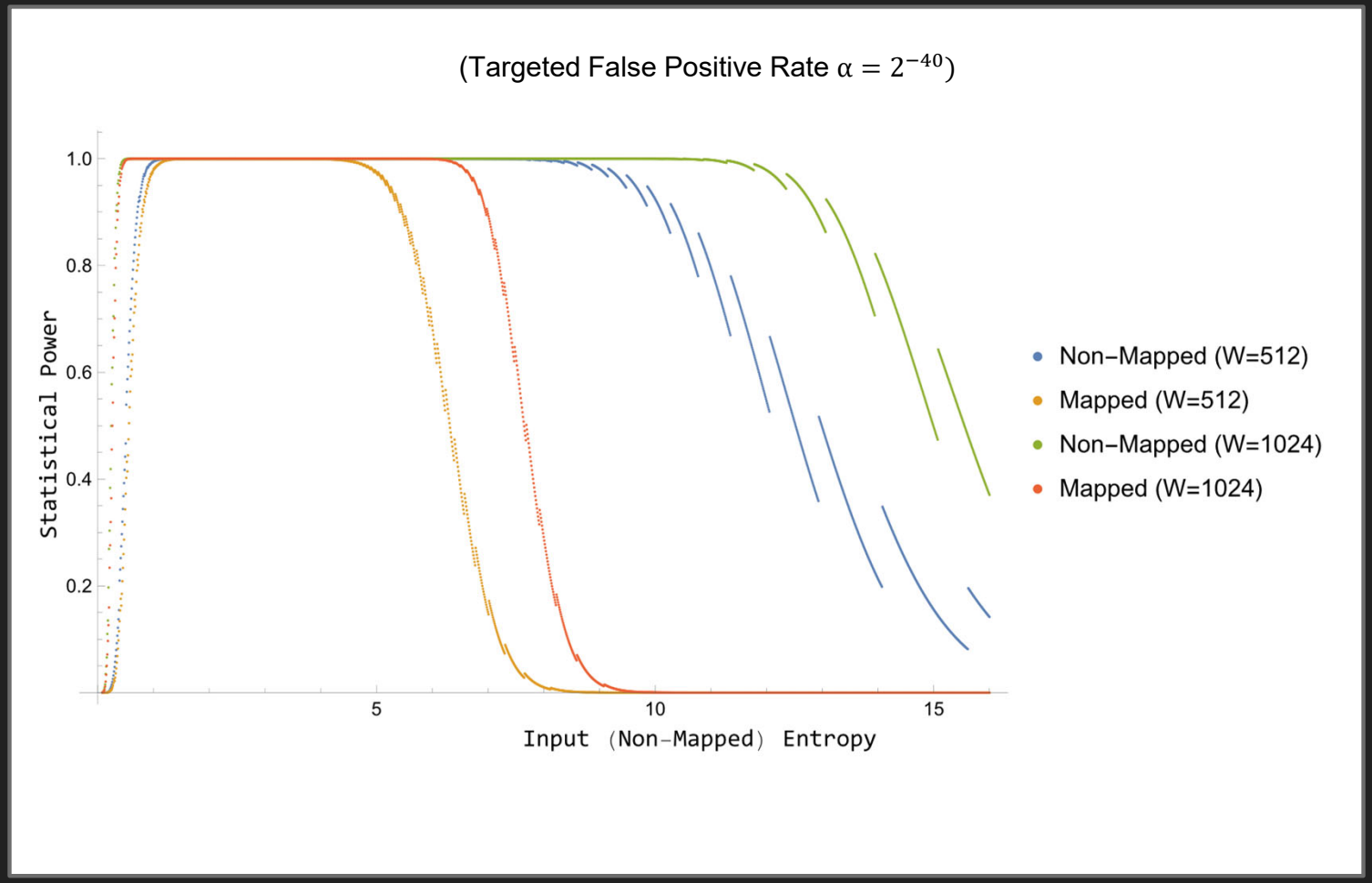


Mapping and the APT: Example Post-Mortem

- In this example, the behavior of the APT using mapped data is dramatically worse than for the APT using unmapped data.
- This occurs across many distributions, so long as there are many values that have a reasonable chance of being output.

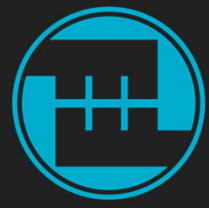


Mapping and the APT: Thousands of Examples

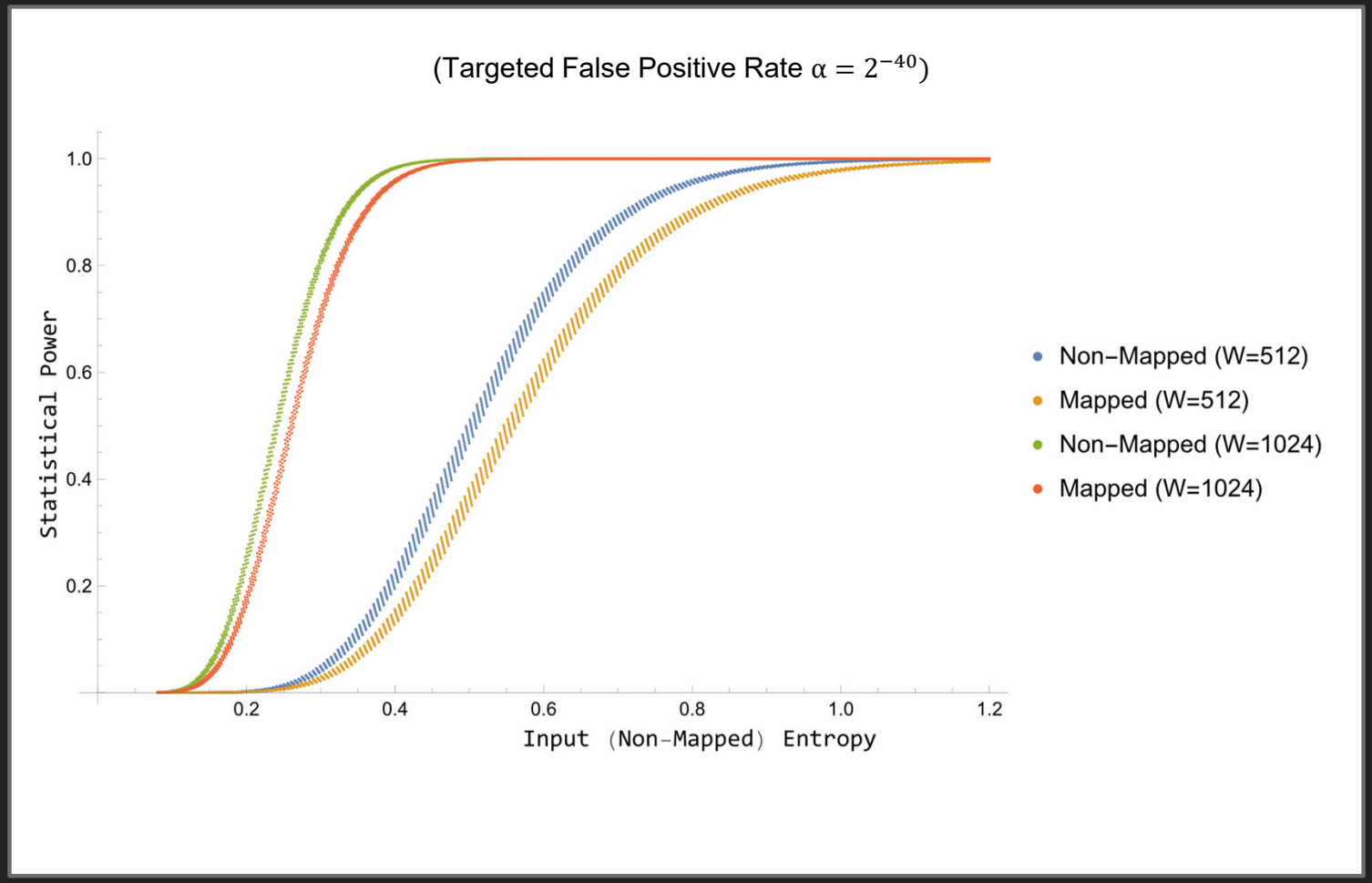


Mapping and the APT: Kilo-Post-Mortem

- For a fixed window size the statistical power for the mapped data is often worse and never better than for the unmapped data.
- For the high-apparent-entropy / “APT cutoff is too small” issue (i.e., the targeted issue for this resolution):
  - The mapping solution proposed by this draft makes the APT dramatically less sensitive for most of the evaluated range.
  - Mapping also undermines the benefits of increasing the window size.

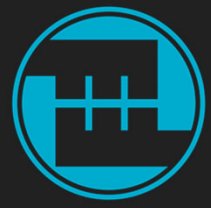


Mapping and the APT: Enhance!



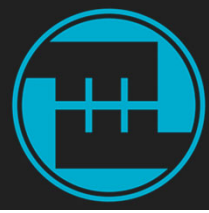
Mapping and the APT: Enhanced-Post-Mortem

- For the low apparent entropy issue (one of the “APT cutoff is too large” issues), increasing the window size helps more than mapping hurts.
- This is because all the other symbols become unlikely when the most likely symbol is very likely to occur.

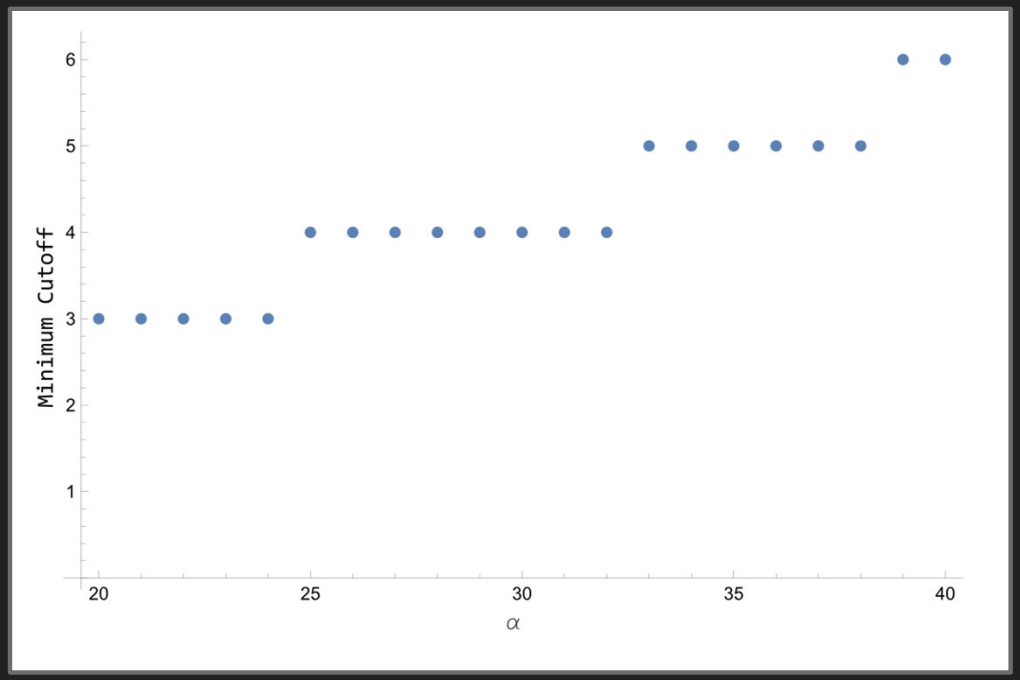


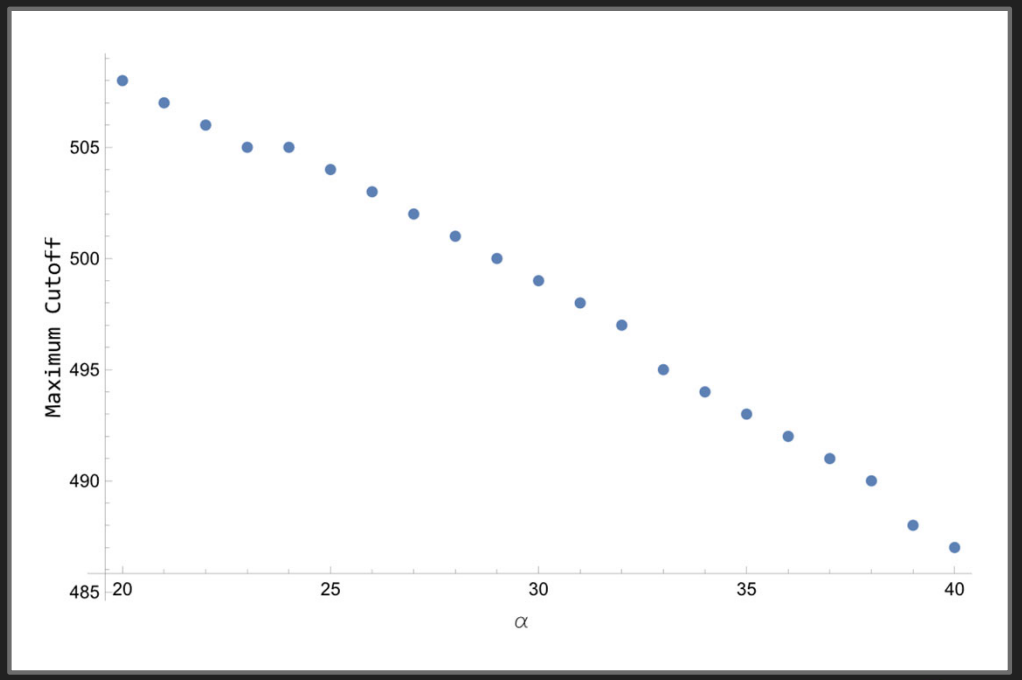
Mapping and the APT: Meta-Post-Mortem

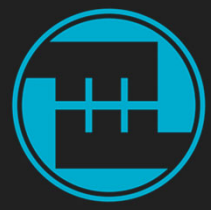
- These examples support the notion that mapping isn't a good general solution to either the "APT cutoff is too small" or "APT cutoff is too large" issues.
- The problem we see here is a lower than desired statistical power when the APT cutoffs are too small or too large.
- In all the tested situations, increasing the window size helps both the "APT cutoff is too large" and the "APT cutoff is too small" issues.
  - The efficacy of the APT tends to increase as the window size increases.
- To satisfy the Criterion (b) requirements using the corrected SP 800-90B Section 4.4.2 cutoff procedure and $W = 512$, cutoffs should be a range that depends on α.



Mapping and the APT: Meta-Post-Mortem







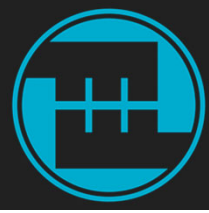
Mapping and the APT: Possible Solution

I suggest that NIST not propose mapping as an approach to resolving low APT statistical power.

- We didn't see any benefit from mapping in either the JEnt Example, or the thousands of bounds calculations.
  - In our examples, some mappings only slightly weakened the APT but no mapping made the test better.
  - It may be possible to construct examples where specific noise source / mapping combinations are helpful, but it isn't a good general-purpose approach.
  - In the absence of careful engineering, mapping is likely to make things worse.

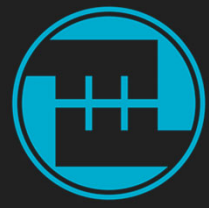
Mapping and the APT: Possible Solution

- The actual relationship between H, α, and β under a specific failure mode is complicated. We can produce cutoff bounds for the smallest recommended α.



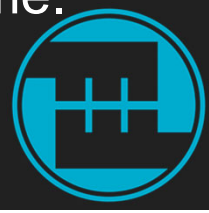
Some comments on the Draft IG D.K [Draft IG D.K, Resolution 22]

- This is, in some sense, a new requirement.
  - Developer-specified health tests have an analogous requirement, but CMVP hasn't been enforcing statistical power requirements for approved health testing in ESV reviews.
  - One could conceptualize this as a specialization of SP 800-90B Section 4.3 Requirement 8 / IG D.K Resolution 14 (Shall IDs #77, #118, and #119, all of which are presently marked optional).



Some comments on the Draft IG D.K [Draft IG D.K, Resolution 22]

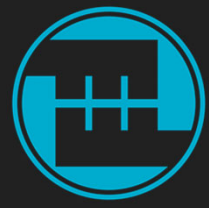
- Irrespective of how this is conceptualized, an explicit transition period seems very important if this change is integrated.
- Any requirements update that necessitates functional changes to entropy source designs is particularly disruptive for hardware vendors.
  - Hardware vendors take years to go from design to sale and they generally cannot make functional changes to implemented designs.
- Prior guidance from NIST suggested that health testing efficacy requirements were a significant goal of the next revision of SP 800-90B.
  - It may make sense to delay integration of these changes until that time.



Proposed Resolution: Option 1

Enforce SP 800-90B Section 4.3 Requirement 8 / IG D.K Resolution 14 (Shall IDs #77, #118, and #119).

- These essentially require that the vendor understand what the failure modes are likely to be and show that the health testing detects these failure modes.
- APT problems would be exposed in this approach.
- This draft IG resolution would not be necessary in this approach.
- This would be quite disruptive for vendors seeking ESV certificates.

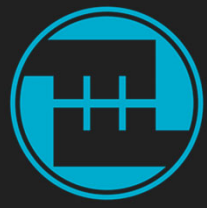


Proposed Resolution: Option 2

Proposed revised text for this resolution:

"If the entropy source uses the Adaptive Proportion Test (APT), the submitter **shall** include a justification that the APT has at least a 50% chance of detecting a known or expected failure mode that results in a min entropy drop from H to H/2 after examining 50000 consecutive samples from the degraded source. If no such justification can be made, then the vendor **shall** include a developer-defined alternative health test that can detect the identified failure mode with at least this probability."

- This is a step toward the meaningful (currently optional) requirements in SP 800-90B Section 4.3 Requirement 8 / IG D.K Resolution 14 (Shall IDs #77, #118, and #119).
- Somewhat disruptive for vendors seeking ESV certificates.

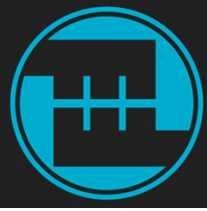


Proposed Resolution: Option 3

Proposed revised text for this resolution:

"If an entropy source uses the Adaptive Proportion Test (APT) with a window size of 512 and an APT cutoff of less than 6 or more than 487, then the laboratory **shall** justify why the APT is meaningful in the Entropy Assessment Report OR the entropy source **shall** use an APT window of at least 1024 symbols (i.e., $W \geq 1024$) and APT cutoffs **shall** be chosen for this larger window size."

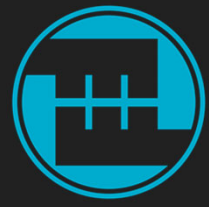
- This restricts the changes to entropy sources that may have problems with low statistical power in the anticipated APT failure mode, proposes a technically useful solution, and allows vendors/laboratories the flexibility to argue that this isn't necessary on a per-entropy source basis.
- This still has some impact on vendors seeking ESV certificates.



Proposed Resolution: Option 4

Do nothing and discard this draft IG resolution.

- The main problems that we identified in this analysis were:
  - The SP 800-90B Section 4.4.2 APT cutoff procedure uses an IID assumption that is invalid for most entropy sources.
  - The APT is designed to detect failure modes that are unlikely to occur in many entropy sources.
  - Very large and very small APT cutoffs yield a statistical power less than that required for developer-defined health tests.
  - A larger window size would be useful for certain low-entropy and high-entropy sources.
- These problems aren't that pressing and can be comprehensively addressed in the next revision of SP 800-90B.



References

- [HD 2012] Patrick Hagerty and Tom Draper. *Entropy Bounds and Statistical Tests*. 2012.
- [HJ 2019] Joshua E. Hill and Benjamin Jackson. *NIST Special Publication 800-90B Comments*. Version 1.9, December 2019.
- [Draft IG D.K] NIST. *D.K Interpretation of SP 800-90B Requirements*. February 6, 2024.