# A Critical Analysis of ISO 17825 ('Testing methods for the mitigation of non-invasive attack classes against cryptographic modules')

Carolyn Whitnall[1] and Elisabeth Oswald[1,2]

[1] University of Bristol, Bristol, UK
[2] University of Klagenfurt, Klagenfurt, Austria
{carolyn.whitnall, elisabeth.oswald}@bristol.ac.uk

**Abstract.** The ISO standardisation of 'Testing methods for the mitigation of non-invasive attack classes against cryptographic modules' (ISO/IEC 17825:2016) specifies the use of the Test Vector Leakage Assessment (TVLA) framework as the sole measure to assess whether or not an implementation of (symmetric) cryptography is vulnerable to differential side-channel attacks. It is the only publicly available standard of this kind, and the first side-channel assessment regime to exclusively rely on a TVLA instantiation.
TVLA essentially specifies statistical leakage detection tests with the aim of removing the burden of having to test against an ever increasing number of attack vectors. It offers the tantalising prospect of 'conformance testing': if a device passes TVLA, then, one is led to hope, the device would be secure against all (first-order) differential side-channel attacks.
In this paper we provide a statistical assessment of the specific instantiation of TVLA in this standard. This task leads us to inquire whether (or not) it is possible to assess the side-channel security of a device via leakage detection (TVLA) only. We find a number of grave issues in the standard and its adaptation of the original TVLA guidelines. We propose some innovations on existing methodologies and finish by giving recommendations for best practice and the responsible reporting of outcomes.

**Keywords:** side-channel analysis, leakage detection, security certification, statistical power analysis

## 1 Introduction

In the late 1990s, Kocher et al. [23] raised awareness of the fact that 'provably secure' cryptography is potentially vulnerable to attacks exploiting auxiliary information not accounted for in traditional security models (e.g. power consumption or other measureable characteristics of devices in operation). Since then, designers and certification bodies have been increasingly concerned with ensuring and evaluating the physical security of cryptographic implementations. Adapting theoretical security models to incorporate the full range of realistic

physical threats is difficult (likely infeasible) [36], so that it is typically considered necessary to subject actual products to experimental testing in a laboratory setting.

The approach taken by testing regimes within the context of Common Criteria (CC) or EMVCo evaluations is to test 'all' of the most effective known attacks developed in the side-channel literature to date (the JHAS group decides on the strategies to be considered). But the growing number of such attacks and the difficulty of determining *a priori* which are the most pertinent to a particular scenario (see e.g. [9,35]) makes this unsustainable. An alternative option could be to rely on *leakage detection* testing along the lines of the Test Vector Leakage Assessment (TVLA) framework first proposed by Cryptography Research, Inc. (now Rambus) [17].

Rather than aim at the successful extraction of sensitive information from side-channel measurements, as an attack-based evaluation would do, leakage detection simply seeks evidence (or convincing lack of evidence) of sensitive data dependencies in the measured traces. TVLA does this via a suite of Welch's $t$-tests targeting mean differences in carefully chosen partitions of trace measurements. For example, the fixed-versus-random test looks for a statistically significant difference between a trace set associated with a fixed plaintext input and another trace set associated with randomly varying inputs. Alternatively, the leakage associated with a specific intermediate value (such as an S-box output) can be targeted by comparing a trace set that has been partitioned into two according to the value of that bit or byte. Both the 'specific' and the 'non-specific' type tests are univariate and are performed on each point in a trace set separately in order to draw conclusions about the overall vulnerability of the implementation. So-called 'higher order' tests exist to target leakage, more complex in its functional form, that does not present via differences in the mean but can be found in higher order (joint) statistical moments; these typically entail pre-processing the traces before performing the same univariate point-wise test procedures [32].

TVLA is the most well-established and widely-adopted suite of leakage detection tests despite the lack of a comprehensive analysis of its performance. Significantly, the ISO standard ISO/IEC 17825:2016 ('Testing methods for the mitigation of non-invasive attack classes against cryptographic modules'; we will refer to it as ISO 17825) [20] specifies TVLA (in its full first-order form, as we describe in Section 2) as the sole required measure for testing against differential side-channel attacks on symmetric key cryptosystems[3]. ISO 17825 ties in with ISO 19790, which is the intended replacement/revision of FIPS 140-2[4] (the

[3] Other detection methodologies exist outside of the TVLA framework (including approaches based on mutual information [6,7,25], correlation [13] and the $F$-statistic [3] – all variants on statistical hypothesis tests, with differing degrees of formalism). These other tests and 'higher order' tests are not part of ISO 17825 and therefore outside the scope of this submission

[4] `https://csrc.nist.gov/Projects/cryptographic-module-validation-program/Standards`

main evaluation scheme in the US). ISO 19790 specifies the much broader goals of a security evaluation, and ISO 17825 focuses on susceptibility to non-invasive attacks for devices aiming for security level 3 or 4.

Within the cryptographic community, publicly available standards are a key mechanism to ensure the widespread adoption of good practice, and we would argue that the same should hold in the area of security evaluations. Yet this is sadly not the case: high-security evaluations according to (e.g.) CC, or EMVCo, do not release the list of threats that JHAS has agreed are relevant for evaluation. Thus ISO 17825 is the only publicly available standard that covers side channel evaluations. As such it is positioned to become *the* standard methodology for side-channel testing outside the existing smart card market (which is dominated by CC and EMVCo). Much is therefore at stake from a commercial as well as an academic perspective when we come to consider how good ISO 17825/TVLA is at the task for which it was designed (conformance testing in the context of side-channel leakage).

We begin this submission by considering the goal(s) of leakage detection in the context of external evaluations generally, followed by some background on TVLA in particular and the relevant ISO standards (see Section 2). We introduce statistical power analysis[5] in Section 3 and, in Sections 4 and 5 use these tools to examine the false positive and false negative error rates implied by the standard recommendations, with appropriate consideration for the fact that multiple tests are performed as part of a single evaluation. We also introduce the notion of coverage, inspired by that of code coverage in software testing, and use this to comment on how thoroughly the recommendations take account of realistic threats. We explore some alternative approaches in Section 6 and conclude with some recommendations for best practice in Section 7. Our analysis is enabled by adapting a novel method for complex statistical power simulations by Porter [27], as well as deriving real-world effect sizes from some actual devices. Interested readers can find more details about statistical power analysis for leakage detection, including in relation to the subtly different goals of *in-house* evaluation, in our companion paper *A Cautionary Note Regarding the Usage of Leakage Detection Tests in Security Evaluation* [42].

## 2 Background: Leakage Detection in a Security Evaluation

Leakage detection is often carried out as part of an exercise to evaluate the security of a cryptographic device. It might be performed by an evaluation laboratory in order to provide security certification when the device goes on sale, or it might be an in-house effort during the development process to highlight and fix potential problems prior to formal external evaluation. We address both scenarios in [42], while here we focus on the context of external evaluations, where there are two potential end results aimed at by a detection test:

[5] 'Power,' as we will explain later in the paper, is a statistical concept and should not be confused with the 'P' of DPA which refers to power consumption.

**Certifying vulnerability:** Find a leak in **at least one** trace point. In such a case it is important to control the number of false positives (that is, concluding there is a leak where there isn't one).

**Certifying security:** Find **no leaks** having tested thoroughly. Here false negatives (failure to find leaks that are really there) become a concern.

As we will see, the statistical methods used for leakage detection cannot 'prove' that there is no effect, they can at best conclude that there is evidence of a leak or that there is no evidence of a leak. Hence it is especially important to design tests with **'statistical power'** in mind – that is, to make sure the sample size is large enough to detect a present effect of a certain size with reasonable probability (see Section 3). Then, in the event that no leak is discovered, these constructed features of the test form the basis of a reasoned interpretation. A further, considerable challenge implicit to this goal is the necessity to be convincingly exhaustive in the range of tests performed – that is, to target 'all possible' intermediates and all relevant higher-order combinations of points. (This suggests analogues with the idea of *coverage* in code testing, which we discuss in Section 5.1).

### 2.1 TVLA and its Adoption Within Standards

The TVLA framework was presented by researchers from Cryptography Research Inc. (now Rambus) at the 2011 Non-Invasive Attack Testing workshop organised by NIST [17]. It describes a series of statistical hypothesis tests to reject (or not) the null of 'no sensitive information leakage' against various alternative hypotheses designed to capture a large range of possible leakage forms and sources. In summary form (see the paper for full details) the procedure is follows:

- An acquisition of size $n$ is taken as the device operates with a fixed key on a fixed plaintext chosen to induce certain values in one of the middle rounds. It is then divided into two disjoint sets FIXED1 and FIXED2, each of size $n/2$.
- An acquisition of size $2n$ is taken as the device operates with the same fixed key on random inputs. It is then divided into two disjoint sets RANDOM1 and RANDOM2, each of size $n$.
- Welch's $t$-tests [41] are performed, with an (implied, for large samples) significance level of $\alpha \approx 0.00001$, comparing the population means of:
  - The fixed-plaintext traces FIXED1 with the random-plaintext traces RANDOM1.
  - The RANDOM1 traces such that a target intermediate takes a certain value, versus the remainder of the RANDOM1 traces, for the following targets: each bit of the XOR between round $R$ input and output; each bit of the $R^{th}$ round SubBytes output; each bit of the round $R$ output; each byte of the round $R$ output (repeated for all possible values in a one-versus-all manner).

– The above is repeated identically for trace sets FIXED2, RANDOM2. The module is considered to fail the overall test if any *pair* of repeated individual tests both conclude that there is a statistically significant difference (in the same direction) at any trace index.

The TVLA specification provides no discussion of the statistical power of this procedure, nor does it explicitly discuss the chosen parameters, nor whether the multiple comparisons problem was accounted for in the design.

### 2.2 ISO Standards for Physical Security

ISO/IEC 19790 [21] specifies four increasingly rigorous security levels and the criteria for achieving them. Levels 3 and 4 require (among other things) that the modules mitigate successfully (to a specified degree) against non-invasive physical attacks including simple power analysis (SPA) and differential power analysis (DPA).

ISO/IEC 17825:2016 [20] specifies the tests that the modules must undergo and the different parameters (sample size, laboratory time, pass/fail criteria) for running the tests according to each security level.

Under this latter standard, the DPA resilience of symmetric key cryptosystems is essentially determined by performing the full suite of first-order TVLA tests as detailed above, with the following main differences:

– Fixed plaintexts are required to have the same special characteristics as the particular values specified by Goodwill et al., but the method of choosing suitable candidates is left up to the analyst.
– The specified risk of false positives (a.k.a. the significance level, typically denoted $\alpha$) is 0.05, which is considerably higher than the level of 0.00001 implied by Goodwill et al.'s $t$-value threshold of 4.5.

Security levels 3 and 4 are separated by the resources available to perform the analysis, and the degree of data pre-processing, as per Tab. 1. These criteria seem to be directly inherited from FIPS 140-2, which originally was based on attacks (like CC and EMVCo evaluations).

The standard leaves ambiguous whether the sample size specifications apply per acquisition or for both fixed and random trace sets combined; similarly whether they are intended per repetition or for both the first and the confirmatory analysis combined. We have assumed fixed and random are counted separately and the two repetitions are counted jointly, so that there are 10,000 or 100,000 each of the fixed input and random input traces, split across the two 'independent' evaluations.

The remaining questions of interest are then how well TVLA, when applied as specified in ISO 17825, succeeds in the goals of certifying vulnerability and/or certifying security – and whether or not (and how) the recommendations could be adapted to do so more effectively. To address these questions we first introduce statistical power analysis, which will give us the tools to analyse (and potentially improve) the theoretical properties of the tests.

| | Level 3 | Level 4 |
|---|---|---|
| Maximum acquisition time per test (hours) | 6 | 24 |
| Maximum overall acquisition time (hours) | 72 | 288 |
| Sample size | 10,000 | 100,000 |
| Synchronisation signal available | Yes | Yes |
| Noise reduction | Averaging (over 10) | Spectrum analysis |
| Static alignment attempted | No | Yes |
| Dynamic alignment attempted | No | Yes? |

**Table 1.** Configuration of the tests to attain security levels 3 and 4. (Note that the overall acquisition time includes tests not related to DPA vulnerability).

## 3 Statistical Power Analysis for Leakage Detection Tests

It is *impossible to eliminate* errors in statistical hypothesis testing; the aim is rather to understand and minimise them. The decision to reject a null hypothesis when it is in fact true is called a Type I error, a.k.a. 'false positive' (e.g. finding leakage when in fact there is none). The acceptable rate of false positives is explicitly set by the analyst at a significance level $\alpha$. A Type II error, a.k.a. 'false negative' is a failure to reject the null when it is in fact false (e.g. failing to find leakage when in reality there is some). The Type II error rate of an hypothesis test is denoted $\beta$ and the **power** of the test is $1 - \beta$, that is, the probability of correctly rejecting a false null in favour of a true alternative. The two errors can be traded-off against one another, and mitigated (but not eliminated) by:

- Increasing the **sample size** $N$, intuitively resulting in more evidence from which to draw a conclusion.
- Increasing the minimum **effect size** of interest $\zeta$, which in our case implies increasing the magnitude of leakage that one would be willing to dismiss as 'negligible'.
- Choosing a different statistical test that is more efficient with respect to the sample size.

For a given test (i.e. leaving aside the latter option) the techniques of **statistical power analysis** are concerned with the mutually determined relationship between $\alpha$, $1 - \beta$, $\zeta$ and $N$. For the simple case of a $t$-test with equal sample sizes and population variances $\sigma_1$ and $\sigma_2$[6], the following formula can be derived (see Appendix A):

$$N = 2 \cdot \frac{(z_{\alpha/2} + z_{\beta})^2 \cdot ({\sigma_1}^2 + {\sigma_2}^2)}{\zeta^2} \tag{1}$$

[6] We consider these conditions to approximately hold in the case of most of the ISO standard tests, where the partitions are determined by uniformly distributed intermediates.

where $\zeta = \mu_1 - \mu_2$ is the true difference in means between the two populations (this relationship can be found in any standard statistics textbook). Note that Eq. (1) can be straightforwardly rearranged to alternatively compute any of the significance level, effect size or power in terms of the other three quantities.

### 3.1 Configuring Tests via an *A Priori* Power Analysis

Ideally, a power analysis is performed before a leakage evaluation takes place as an aid to experimental design; this is known as *a priori* power analysis and can help to ensure (e.g.) the collection of a large enough sample to detect data-dependencies of the expected magnitude with the desired probability of success [25]. Power analysis can be performed *after* data collection in order to make statements about the power to detect a particular effect size of interest, or the minimum effect size that the test would be able to detect with a certain power. This can be useful when it comes to responsibly interpreting the non-rejection of a null hypothesis. However, it is crucial that the effect sizes are chosen independently of the test, based on external criteria, as it has been shown that attempts to estimate 'true' effect sizes from the test data produce circular reasoning. In fact, there is a direct correspondence between the $p$-value and the power to detect the observed effect, so that 'post hoc power analysis' merely re-expresses the information contained already in the test outcome [18].

Also needed in order to perform statistical power analysis are the population standard deviations of the partitioned samples, which may or may not be the same. These are usually assumed to have been obtained from previous experiments and/or already-published results, which can be especially tricky when approaching a new target for evaluation.

### 3.2 Effect Size

This requirement for information *about* the data sample which cannot be estimated *from* the data sample is the main obstacle to statistical power analysis. The choice of effect sizes for the computations can be guided by previous experiments (e.g., in our case, leakage evaluation on a similar device with a similar measurement set up) or (ideally) by some rationale about the practical implications of a given magnitude (e.g. in terms of loss of security). Note that we always eventually need some rationale of this latter type: what is ultimately of interest is not just whether we are able to detect effects but whether the effects that we detect are of practical concern. With a large enough sample we will always be able to find 'arbitrarily small' differences; the question then remains, at what threshold do they *become* 'arbitrary'?

It is convenient (and bypasses some of the reliance on prior information) to express effect sizes in standardised form. Cohen's $d$ is defined as the mean difference divided by the pooled standard deviation of two samples of (univariate) random variables $A$ and $B$:

$$d = \frac{\overline{a} - \overline{b}}{\sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}}$$

where $\overline{a}$, $\overline{b}$ are the sample means, $s_A^2$, $s_B^2$ are the sample variances and $n_A$, $n_B$ are the sample sizes. Notice that this is essentially a measure of signal-to-noise ratio (SNR), closely related to (and therefore tracking) the various notions that already appear in the side-channel literature. The formula for the sample size required for the $t$-test can be expressed in terms of the standardised effect size as follows:

$$N = 4 \cdot \frac{(z_{\alpha/2} + z_{\beta})^2}{d^2} \tag{2}$$

Cohen [8] proposed that effects of 0.2 or less should be considered 'small', effects around 0.5 are 'medium', and effects of 0.8 or more are 'large'. Sawilowsky [30] expanded the list to incorporate 'very small' effects of 0.01 or less, and 'very large' and 'huge' effects of over 1.2 or 2.0 respectively. The relative cheapness of sampling leakage traces (and subsequent large sample sizes) compared with studies in other fields (such as medicine, psychology and econometrics), as well as the high security stakes of side-channel analysis, make 'very small' effects of more interest than they typically are in other statistical applications.

Focusing on standardised effects helps to put the analysis on a like-for-like footing for all implementations, but it doesn't remove the need for specific knowledge about a device in order for meaningful interpretation.

### 3.3 The Impact of Multiple Testing

Statistical hypothesis testing is generally introduced under the implicit assumption that a single null/alternative pair is up for consideration. Unfortunately, controlling error rates becomes even more complicated when multiple tests are performed as part of the same experiment. Without appropriate modifications, test conclusions are no longer formally supported. This is because, if each test has (by design) a probability $\alpha$ of falsely rejecting the null hypothesis, then the probability of rejecting *at least one* true null hypothesis across all $m$ tests (that is, the overall false positive rate as opposed to the per-test rate) might be as high as $\alpha_{\text{overall}} = 1 - (1 - \alpha_{\text{per-test}})^m$ if those tests are independent. (Otherwise, the rate will be lower but will depend on the form of the dependencies).

**Multiplicity Corrections** In the statistics literature there are two main approaches to correcting for multiple tests: controlling the *family-wise error rate* (FWER) and controlling the *false discovery rate* (FDR). Both of these were discussed and evaluated in the context of leakage detection by Mather *et al.* [25].

FWER-based methods work by adjusting the per-test significance criteria in such a way that the *overall* rate of Type I errors is no greater than the desired $\alpha$ level. For example:

- Bonferroni correction [12]: per-test significance level obtained by dividing the desired overall significance level by the number of tests $m$, i.e. $\alpha_{\text{per-test}} = \frac{\alpha}{m}$. Controls the FWER for the 'worst case' scenario that the tests are independent, and is conservative otherwise.

– Šidák correction [40]: explicitly *assumes* independence, and that all null hypotheses are false, and sets $\alpha_{\text{per-test}} = 1 - (1 - \alpha)^{\frac{1}{m}}$. These assumptions potentially gain power but are unlikely to suit a leakage evaluation setting.
– Holm adjustment [19]: a 'step up' procedure; tests are ordered according to $p$-value (smallest to largest), and criteria set such that $\alpha_i = \frac{\alpha}{m-i+1}$ for the $i^{th}$ test.

It should be clear that any such downward adjustment to the per-test Type I error rates (i.e. in order to prevent concluding that there is a leak when there isn't) inevitably increases the rate of Type II errors (the probability of missing a leak which is present). Erring on the "safe side" with respect to the former criterion may not be at all "safe" in terms of the cost to the latter. The relative undesirability of the two error types depends heavily on the application and must be carefully considered.

FDR-based methods take a slightly different approach which is more relaxed with respect to Type I errors and subsequently less prone to Type II errors. Rather than minimise the probability of *any* false positives they instead seek to bound the proportion of total 'discoveries' (i.e. rejected nulls) which are false positives. The main FDR-controlling method, and the one that we will consider in the following, is the Benjamini–Hochberg procedure, which (like the Holm correction) operates in a 'step up' manner as follows:

1. For the ordered (small to large) $p$-values $p_{(1)}, \ldots, p_{(m)}$, find the largest $k$ such that $p_{(k)} \leq \frac{k}{m}\alpha$.
2. Reject the null hypothesis for all tests $i = 1, \ldots, k$.

A recent proposal in the side-channel literature [11] takes an alternative third way, using methods developed for the purpose of performing a meta-analysis based on multiple independent studies: the decision to collectively reject or not reject a set of null hypotheses is based on the *distribution* of the $p$-values. (We do not analyse this method in the following due to its heavy reliance on the independence assumption).

In addition to the inevitable loss of power associated with all of the above adjustments, a substantial obstacle to their use is the difficulty of analysing (and controlling) the power, which is essential if we want to draw meaningful and comparable conclusions from test outcomes. In cases where a single per-test significance level $\alpha_{\text{per-test}}$ is derived (e.g. Bonferroni and Šidák), this can simply be substituted into the power analysis formulae to gain the per-test power. However, consensus is lacking when it comes to performing equivalent computations for FDR-controlling procedures (compare, e.g., [4,14,24,28,39]; in Section 6 we adopt an approach by Porter that operates by simulating test statistics but is constrained to fully specified test scenarios [27]). Moreover, depending on the over-arching goal of the analysis, per-test power may not even be the relevant quantity to consider, as we next discuss.

**Different Notions of Power** Just as multiple tests raise the notion of an 'overall' Type I error rate which is not equal to the per-test error rate, so it is

worth giving thought to the 'overall' Type II error and what precisely we mean by that. We have seen above that multiplicity corrections reduce the per-test power – the probability of detecting a true effect wherever one exists. Porter [27] describes this as 'individual' power, and contrasts it with the notion of '$r$-minimal' power[7] – the probability of detecting at least $r$ true effects. We propose that the 1-minimal power is the relevant notion in the context of certifying vulnerability/security, since a single detected leak is sufficient to fail a device.

The probability of detecting *all* true effects (as might be the goal of an in-house development-time evaluation) is known as the 'complete power'. The $r$-minimal power is naturally greater than or equal to this quantity. In particular, the 1-minimal power can actually be *higher* in a multiple testing scenario than in a single test – as long as the true number of false positives is greater than 1, each such test represents an additional opportunity to find an effect.

## 4 ISO 17825 for Certifying Vulnerability

In this section we examine how reliable ISO 17825 is for certifying vulnerability – demonstrating a sensitive dependency in the trace measurements. Since a single significant test outcome is sufficient to fail the device, it is crucial that the probability of a false positive be kept very low.

Under the standard, the per-test rate is controlled at $\alpha_{\text{per-test}} = 0.05$ (see Subsection 11.1), and no adjustment is made for the fact that each test is performed against multiple (potentially thousands of) trace points. However, any discovered vulnerability is required to be confirmed by a second test on a separate, identically acquired dataset. In either one of the two sets of tests we would expect that (on average, under the assumption of independence) 5 in every hundred true null hypotheses will be falsely rejected, so that for long traces the overall probability of a false detection becomes almost one. The probability of *both* sets of tests producing a false positive is $(1 - (1 - \alpha_{\text{per-test}})^m)^2$; the probability of this happening such that the sign of both the effects is the same is $(1 - (1 - \alpha_{\text{per-test}})^m) \times (1 - (1 - \alpha_{\text{per-test}}/2)^m)$ (the product of an error of any direction in the first test and an error of fixed direction in the second; see the red lines in Fig. 1). However, the probability of observing two false positives (of the same sign) *in the same position* is $\alpha_{repeat} = 1 - \left(1 - \frac{\alpha^2_{\text{per-test}}}{2}\right)^m$, which grows much slower as $m$ increases (see the yellow lines in Figure 1). Still, under the standard-recommended significance criterion of $\alpha_{\text{per-test}} = 0.05$, the probability of at least one coinciding detection is over a half once the length of the trace reaches 600. By contrast, under the original TVLA recommendations (which imply $\alpha_{\text{per-test}} \approx 0.00001$), the probability of a coinciding detection is close to zero even for traces that are millions of points long. (Only once the number of points is on the order of $10^{10}$ do coinciding false detections become non-negligibly probable).

[7] Porter uses the terminology $d$-minimal; we use $r$ instead of $d$ to avoid confusion with Cohen's $d$.

The standard fails to provide adequate assurance that detected vulnerabilities are real unless leakage traces are extremely short. Either a stricter per-test significance criterion (combined with the repetition step) or an established method to control the FWER (see the purple lines in Figure 1) would be preferable for this purpose.

The probability of a false detection under an FDR-controlling procedure depends on the density of true leaks within the trace and is less easy to state in advance in this way; note however that such methods do not *claim* to avoid false detections altogether, rather to ensure that they are few relative to the number of true effects identified. We provide some analysis in Section 6, essentially confirming that they are ill-suited to the goal of certifying vulnerability, where a single false positive is enough to fail a device altogether according to the standard.

The question of how best to handle multiple comparisons depends not just on the ability of each option to avoid false positives but on the power of each to detect true positives (i.e. their ability to avoid false negatives). We address this within the next section, as we turn our attention to the standard's capabilities when it comes to certifying security.

## 5 ISO 17825 for Certifying Security

We have argued so far that the discovery of a leak when the standard recommendations are followed does not reliably certify vulnerability, due to the high risk of a false positive. We now ask the complementary question: what, if anything, can be concluded if a leak is *not* detected? Can non-discovery be interpreted to 'certify security'?

This question is best separated into two: have all realistic vulnerabilities been tested for? and can we trust the conclusions of all the tests that were performed? The first of these is the simpler to answer.

### 5.1 Have All Realistic Vulnerabilities Been Tested For?

In code testing, the extent to which everything that *could* be tested *has* been tested is referred to as 'coverage' [26]. Typical metrics in this setting include

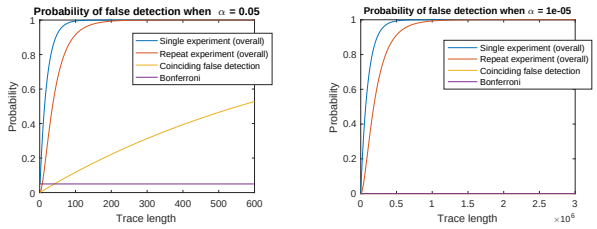


**Fig. 1.** Overall probability of a false positive as the length of the trace increases, for two different per-test significance levels.

code coverage (have all lines of code been touched by the test procedure?), function coverage (has each function been reached?), and branch coverage (have all branches been executed?) [1]. In a hardware setting one might alternatively (or additionally) test for toggle coverage (have all binary nodes in the circuit been switched?) [37]. These examples all assume white-box access to the source code; in black-box testing scenarios, coverage might alternatively be defined in functional terms.

We suggest that the concept of coverage is a useful one for thinking about the (in)adequacy of a side-channel evaluation. The types of questions we might consider include:

- Have all possible intermediates been tested?
- Have all possible leakage forms been taken into account? For example, some circuits might leak in function of the intermediate values; some in function of the *transitions between* certain intermediates; some in combination of both. Differences might present in distribution means or more subtly, such as in higher order moments (e.g. in the presence of countermeasures).
- Have all possible locations in the trace been tested (with each intermediate and leakage form in mind)? This includes all relevant *tuples* of trace points in the case where higher order leakage of protected intermediates is of concern.
- What proportion of the input space has been sampled? Some key/input combinations might be more 'leaky' than others; with a total possible input space of, e.g. (in the case of AES-128) $2^{128} \times 2^{128} = 2^{256}$ (key, plaintext) pairs, it is unavoidable that one can only test a tiny fraction, and we are typically obliged to rely on simplifying assumptions (e.g. 'Equal Images under different Subkeys (EIS)' [31]) in order to interpret outcomes as representative.
- Have all possible side-channels been tested?! With most of the literature typically focused on power (and sometimes EM radiation [16,29]) it is easy to forget that other potentially exploitable characteristics (timing [22], temperature [5], light [15,34] and sound [2,33] emissions) can also be observed.

It should be clear from the description in Section 2.2 that the coverage of ISO 17825 is quite limited. It considers first-order univariate leakages only, relies on one fixed key and (in the case of the fixed-versus-random tests) one fixed input to be representative of the entire sample space, and is confined to a small number of target values (although the fixed-versus-random tests do aim at non-specific leakages). Moreover, by relying solely on the $t$-test the evaluations are only able to discover differences that exhibit in the *means* of the partitioned populations – more general distributional differences (such as those produced by masking in parallel) will remain completely undetected.

### 5.2 How Reliably do the Performed Tests Find Leakage?

Formally, a statistical hypothesis test either rejects the null hypothesis in favour of the alternative, or it 'fails to reject' the null hypothesis. It does not 'prove' nor even 'accept' the null hypothesis. Moreover, it does this with a certain probability of error.

Whilst the Type I error rate $\alpha$ is provided by the standard (albeit chosen badly), the Type II error rate (denoted $\beta$) – i.e. concluding that there is no leak when there is – is opaque to the evaluator without further effort. If this rate is very high (equivalently, we say that the 'statistical power' $1 - \beta$ is low) then the failure of the test to detect leakage really doesn't mean very much at all.

So, if a test fails to reject the null of 'no leakage' in the context of an evaluation, we must be able to say something about its power. The ability of a device to withstand a well-designed test which is known to be powerful indicates far more about its security than its ability to withstand an ad-hoc test which may or may not be suitable for purpose. In addition, the more the statistical properties of the applied methodologies are known and managed, the easier it becomes to compare evaluations across different targets and measurement set-ups, and to establish criteria for fairness. We therefore turn to the tools of statistical power analysis.

Recall from Section 3 that the power of a test depends on the sample size, the standardised effect size of interest (alternatively, the raw effect size and the variance of the data), and the significance criteria (the pre-chosen rate of Type I errors). The standard specifies sample sizes of 10,000 and 100,000 for each of the security levels 3 and 4 respectively, and an (unadjusted) per-test significance criteria of $\alpha_{\text{per-test}} = 0.05$. The actual effect size (if an effect exists) is necessarily unknown (if it was known the evaluator wouldn't need to test for its existence) and depends on the target implementation even if a perfect measurement set-up were available. But we *can* answer the following:

- What is the power of the tests (as specified) to detect the standardised effects as categorised by Cohen and Sawilowsky?
- What effect sizes can the tests (as specified) detect for a given power (for example, if the analyst wishes to balance the rates of the two types of error)?
- What effect sizes have been observed in practice, and would the current specifications need to be revised in order to detect these?

**Power of a Single Test** The LHS of Tab. 2 shows that, of the standardised effects as categorised by Cohen and Sawilowsky, all but the 'very small' are detected with high probability under the sample size criteria defined by the standard. Meanwhile, level 3 and 4 criteria are both inadequate to detect standardised effects of 0.01. (Remember though that a single test essentially corresponds to a leakage trace of unrealistic length 1).

The RHS of the table shows the effect sizes that *are* detectable; for example, an analyst who wishes to control Type II errors at the same rate as Type I errors ($\beta = \alpha = 0.05$) is able to detect effects of size 0.072 under the level 3 criteria and 0.023 under the level 4 criteria. By comparison, the minimum detectable effect sizes for balanced error rates are more than doubled under the original TVLA significance criterion (which approximates to $\alpha = 0.00001$): 0.174 with a sample size of 10,000 and 0.055 with a sample size of 100,000. (See Tab. 6 in App. B).

A natural next question is what size *are* the effects exhibited in actual trace acquisitions, and are the criteria laid out in the standard adequate to detect real-

| Cohen's $d$ | Power | |
|---|---|---|
| | Level 3 | Level 4 |
| Very small (0.01) | 0.072 | 0.352 |
| Small (0.2) | 1.000 | 1.000 |
| Medium (0.5) | 1.000 | 1.000 |
| Large (0.8) | 1.000 | 1.000 |
| Very large (1.2) | 1.000 | 1.000 |
| Huge (2) | 1.000 | 1.000 |

| Power | Cohen's $d$ | |
|---|---|---|
| | Level 3 | Level 4 |
| 0.75 | 0.053 | 0.017 |
| 0.80 | 0.056 | 0.018 |
| 0.90 | 0.065 | 0.021 |
| 0.95 | 0.072 | 0.023 |
| 0.99 | 0.086 | 0.027 |
| 0.99999 | 0.124 | 0.039 |

**Table 2.** LHS: Power to detect Cohen's and Sawilowsky's standardised effects under the level 3 ($N = 10,000$) and level 4 ($N = 100,000$) criteria; RHS: Minimum effect sizes detectable for increasing power thresholds, under the level 3 ($N = 10,000$) and level 4 ($N = 100,000$) criteria.

world vulnerabilities? We seek indicative answers via analysis of some example scenarios.

**Observed Effect Sizes from Realistic Devices** It is not straightforward to 'simply' observe magnitudes in existing acquisitions; all estimated differences will be non-zero, and deciding which ones are 'meaningful' essentially corresponds to the task of detection itself. Choosing 'real' effects based on the outcomes of $t$-tests, and then using the magnitudes of those effects to make claims about 'detectable' effect sizes, amounts to circular reasoning, and depends on the choice of significance criteria. Fortunately the motivation behind leakage detection provides us with a natural, slightly more objective, criterion for identifying 'real' effects, via the outcomes of key recovery attacks. That is, if leakage detection is geared towards identifying (without having to perform attacks) points in the trace which are vulnerable to attack, then an effect size which is 'large enough' to be of interest is one that can be successfully exploited.

We take this approach, and perform distance-of-means attacks on all 128 bits of the first round SubBytes output for three AES acquisitions, taken on an ARM-M0 processor, an 8051 microcontroller and an RFID (i.e. custom ASIC) device. We also compute the sample effects for each of those bits, which enables us to report estimated effect sizes of interest.

To mitigate for false positives we (adapting from [38]) take measures to confirm the stability of an outcome before classifying a point as 'interesting': we repeat the attack on 99% of the full sample and retain only those points where the correct subkey is ranked first in both instances.

Figure 2 shows the raw (top) and standardised (bottom) observed effect sizes (i.e. mean differences associated with an S-box bit) of first round AES traces measured from an ARM-M0 processor, an 8051 microcontroller and an RFID (custom ASIC) device respectively. As expected, because of the different scales of the measurements (arising from different pre-processing, etc), the raw effects are not necessarily useful to compare. The ARM effects range up to about 0.8, while effects on the 8051 and the RFID implementation range up to 3 and 2

respectively. The standardised effects are much more comparable ($\approx 0.6$ and $\approx 1$ for ARM and 8051 respectively; $\approx 0.4$ for the RFID, although this is for the second rather than the first S-box as the latter is less 'leaky' in this instance). [8]

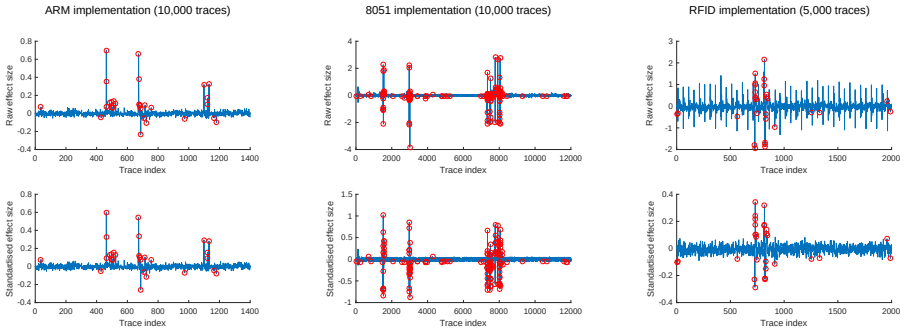


**Fig. 2.** Difference of means (top) and standardised equivalent (bottom) associated with the first bit of the first S-box of two software AES implementations and the first bit of the second S-box of one hardware implementation. Red circles denote points where a distance-of-means attack achieves stable key recovery.

Tab. 3 summarises the standardised and raw effect sizes associated with distance-of-means key recoveries over *all* bits of all S-boxes. The smallest standardised effect detected is 0.0413 for the 8051 microcontroller; the ARM and RFID smallest effects are in a similar ballpark.

| **Implementation** | **Proportion** | **Standardised** | | | **Raw** | | |
|---|---|---|---|---|---|---|---|
| | **interesting** | **Min** | **Max** | **Median** | **Min** | **Max** | **Median** |
| ARM | 0.0226 | 0.0444 | 0.9087 | 0.1155 | 0.0388 | 1.0265 | 0.1073 |
| 8051 | 0.0150 | 0.0413 | 1.4265 | 0.1670 | 0.0254 | 5.3808 | 0.1469 |
| RFID | 0.0049 | 0.0624 | 0.3935 | 0.0933 | 0.2272 | 3.4075 | 0.3836 |

**Table 3.** Summary of effect magnitudes associated with stable distance-of-means key recovery attacks.

[8] In a non-specific fixed-versus-random experiment (even more so in a fixed-versus-fixed one) the differences depend on more than a single bit so, depending on the value of a given intermediate under the fixed input, can potentially be several times larger (see e.g. [32]) – or they can be smaller (e.g. if the leakage of the fixed intermediate coincides with the average case, such as the (decimal) value 15 in an approximately Hamming weight leakage scenario). It is typically assumed in the non-specific case that, as the input propagates through the algorithm, at least some of the intermediates will correspond to large (efficiently detected) class differences [13].

Taking 0.04 as an indicative standardised effect size for actual trace measurements would lead us to conclude that the level 4 criterion is adequate if the full sample of size 100,000 is used in an individual (non-repeated) test, but that the level 3 criterion of 10,000 is not. Using the sample size formula we obtain that a minimum of 32,487 traces are needed to detect an effect of size 0.04 in a single test with balanced error rates $\alpha = \beta = 0.05$. (In reality, one type of error may be deemed more or less of a concern than the other; we state results for balanced rates merely by way of example).

However, data-intensive research has been carried out into the exploitable leakage of devices with far less 'neat' side-channel characteristics than the (comparatively) favourable scenarios exampled above. De Cnudde et al. [10], for example, perform successful attacks against masked hardware implementations with up to 500 million traces, implying both that *extremely* small effects exist and that researchers (and, presumably, some 'worst case' attackers) have the resources and determination to detect and exploit them. FIPS 140-2 (and thus ISO 19790) was conceived to be more economic than CC, but this comes at the cost of not being adequate for state of the art hardware implementations.

We would argue that effects of real world relevance should be extended to include a new category: 'tiny' effects of standardised size $d = 0.001$. An evaluation with $\alpha = 0.05$ and a sample of size of 10,000 or 100,000 (as per the levels 3 and 4 criteria respectively) would have power of just 0.028 or 0.036 respectively to detect such an effect. To achieve a power of 0.95 (that is, balanced error rates) would require a sample of size nearly 52,000,000. Clearly, leakage of this nature is beyond the scope of the ISO standard to detect, whilst still representing a demonstrably exploitable vulnerability.

Furthermore, in practice, of course, evaluators are *not* just checking for a single effect via a single test, but for a range of different effects all in a series of separate (possibly correlated) trace points. This adds considerably to the challenge of rigorous and convincing analysis, due to the problem of multiple comparisons discussed above – corrections for which inevitably impact on the power.

**'Overall' Power in an Example Scenario** The per-test power can be computed via the formulae in Section 3, but the $r$-minimal and the complete power of a set of tests depends on the total number of tests and the ratio of true to false null hypotheses, as well as the covariance structure of the test statistics. This information is not available if an evaluation is set up according to ISO 17825 (it would need to be determined in preliminary experiments).

By way of illustrative analysis we consider the scenario described above in Section 5.2, where there appeared to be around 30 true leak points in a (truncated) first round AES software trace of length 1,400, and we make the simplifying assumption that the tests are independent (we will relax this in Section 6 and show that it makes little difference).

Tab. 4 shows the per-test and the 1-minimal power under the standard specifications to detect two different effect sizes: the empirically observed effect of

standardised size 0.04, and the 'worst case adversary' inspired 'tiny' effect of 0.001. The level 3 sample size is just short of that required to achieve an overall (i.e. 1-minimal) power of $1 - \alpha$ to detect at least one effect of the observed size when the repetition is performed[9]; the level 4 sample size detects it with high probability (even at the stricter TVLA-recommended $\alpha$-level, see Tab. 7 in App. B); however, to detect the 'tiny' effect would require 170 times as many measurements (1,700 more for $\alpha = 0.00001$). Thus, for this scenario at least (and under our simplifying assumptions) we conclude that the standard recommendations are adequate to certify security with respect to modest effect sizes.

Recall, though, that the standard recommendations are inadequate to certify vulnerability, as the overall false positive rates are considerably higher than should be tolerated by a procedure that fails a device based on a single rejected null hypothesis (see Section 4)–this is a prime example that error rates can be 'traded off'. The question is therefore whether any set of parameters or alternative method for multiplicity correction is able to make a better trade-off between the overall false negative and false positive rates.

| **Effect** | **Repeat test?** | **Level 3** | | **Level 4** | | **Required sample size** | |
|---|---|---|---|---|---|---|---|
| | | **Ave** | **1-min** | **Ave** | **1-min** | **Ave** | **1-min** |
| 0.04 | No | 0.516 | 1.000 | 1.000 | 1.000 | 32,487 | 1,055 |
| 0.04 | Yes | 0.086 | 0.932 | 0.988 | 1.000 | 76,615 | 10,647 |
| 0.001 | No | 0.028 | 0.574 | 0.036 | 0.665 | 51,978,840 | 1,687,843 |
| 0.001 | Yes | 0.001 | 0.022 | 0.001 | 0.031 | 122,584,748 | 17,034,581 |

**Table 4.** Average ('per-test') and 1-minimal ('overall') power to detect observed and 'tiny' effect sizes under the level 3 and 4 criteria, and the sample size required to achieve balanced errors for a significance criterion of $\alpha = 0.05$. (30 leak points in a trace set of length 1,400).

## 6 Exploring Alternative Test Configurations

We wish to extend the analysis above to a wider range of adjustment methods in order to see if any emerge as being promising alternatives to the current recommendations. Porter suggests a way to approximate the different types of power by simulating large numbers of test statistics under a suitable alternative hypothesis, performing the multiplicity adjustments and simply counting the proportion of instances where 1, $r$, or all the false nulls are rejected (for the 1-, $r$-minimal and complete powers) as well as the total proportion of false nulls rejected (for the average individual power) [27]. An advantage of this approach

[9] We compute the per-test power under the repetition step as the square of the power to detect with half the sample, deriving from the assumption that the two iterations of the test are independent.

is that it also allows us to relax the independence assumptions underpinning the computations in Tab. 4 – but this introduces the considerable limitation that specific and detailed information about the particular leakage scenario is needed. In a real evaluation we do not typically have this; however, for the purposes of illustration we take the dataset analysed in Section 5.2 as an example scenario from which to construct a realistic set of null and alternative hypotheses, with the aim of showing how the different notions of power evolve as the sample size increases.

Suppose the $t$-statistics corresponding to a trace set of length 1,400 have the same correlation structure as the observed ARM traces, characterised by the covariance matrix $\Sigma$. The null hypothesis is that none of the points leak; the alternative is that there are 30 effects of standardised size 0.04, located as per the analysis presented in Figure 2, where $\mathcal{T}$ denotes the set of indices of successful attacks. Under the null hypothesis, for a large enough trace set (which we need anyway to detect such a small effect) the joint distribution of the $t$-statistics under the alternative hypothesis can be approximated by a multivariate normal with mean $\mu = [\mu_1, \ldots, \mu_{1400}]$ such that $\mu_t = 0.04$ for all $t \in \mathcal{T}$ and $\mu_t = 0$ for all $t \notin \mathcal{T}$, and covariance matrix $\Sigma$. By drawing repeatedly from this distribution and noting which of the (individual) tests, with and without correction, reject the null hypothesis and which do not, we can estimate the power and the error rates for tests in this particular scenario.

We performed the analysis for two different significance levels ($\alpha_{ISO} = 0.05$ and $\alpha_{TVLA} = 0.00001$) and six different methods: no correction, Bonferroni, Šidák and Holm corrections to control the FWER, the Benjamini–Hochberg procedure to control the FDR, and the experiment repetition (for a given overall sample size) as per ISO and TVLA recommendations. Figure 3 shows, for $\alpha_{ISO} = 0.05$, what we consider to be the most relevant results, based on 5,000 random draws from the distribution under the alternative hypothesis. (In particular, the three FWER-controlling corrections perform near-identically, and so we only display a single representative). Figure 6 in App. B shows the corresponding results for $\alpha_{TVLA} = 0.00001$.

It is clear that the different approaches have substantially different characteristics in practice. The FWER-controlling procedures, represented by Bonferroni, successfully keep false positives down at only a small cost to the power relative to the repetition step. The FDR-controlling procedure, meanwhile, has better power than the repetition step but a comparable false positive rate as the sample size increases. At the lower $\alpha$ level implied by the TVLA criteria Bonferroni (as well as the BH procedure) actually has higher power than the repetition step, and all methods keep false positives low for the (short) trace length in question. Moreover, they all achieve high probability of detecting at least one of the 30 leaks within the level 4 sample size threshold.

We repeated the experiment assuming independence between the tests, and found that it made very little difference to either error rate. This is *not* to say that *taking the dependence structure into account in the tests themselves* would not improve the performance of the tests, but it does imply that (at least in

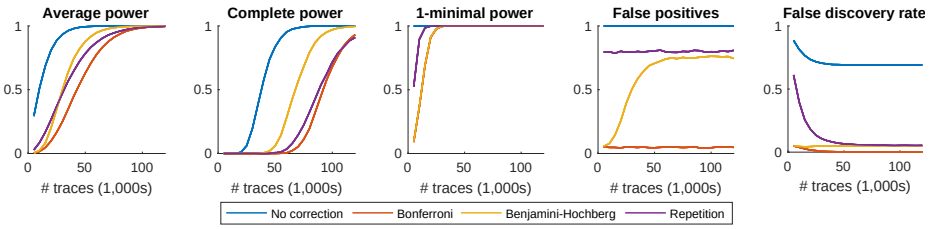


**Fig. 3.** Different types of power and error to detect 30 true effects of size 0.04 in a trace set of length 1,400, as sample size increases, for an overall significance level of $\alpha = 0.05$. (Based on 5,000 random draws from the multivariate test statistic distribution under the alternative hypothesis).

this instance) a power analysis which assumes independence need not give a misleading account of the capabilities of the chosen tests.

In this example scenario, then, the FWER controlling procedures (but not the FDR controlling one) appear favourable to the ISO standard confirmation requirement, holding all other parameters of the ISO standard fixed. However, we have not yet fully explored the impact of the *length* of the trace on their performance, and many real-world evaluations involve considerably more tests than the 1,400 we here consider. Porter's methodology does not readily scale – and, besides, requires specifying a covariance structure. Instead, then, given the similarity of our results under the independence assumption, we proceed on that simplifying basis and take advantage of the fact that the Bonferroni-corrected tests (by contrast with the BH procedure, which we have already been able to rule out) are relatively straightforward to examine analytically.

The obstacle remains, though, that *overall* notions of power – such as 1-minimal, which we have argued is the relevant quantity for our purposes – will always be highly dependent on the (*a priori* unknown) particulars of the evaluation scenario under consideration. In particular, if a longer trace implies more leakage points, then the increased opportunity to detect leakage might help to compensate for the stricter criteria enforced by the Bonferroni procedure (and similar). On the other hand, if the number of leakage points stays fixed as the trace length increases, there is no compensation for the loss of per-test power. We therefore consider a range of hypothetical scenarios: fixed leakage density of 1 in 1,000 and 1 in 100 as the trace length increases; fixed number of leaks at 1 and (as per our example scenario) 30 as the trace length increases. (In the latter, we suppose that the first 30 trace points are the vulnerable ones and all those subsequently added are random).

Figure 4 presents the FWER and the 1-minimal ('overall') power of the unadjusted, repeated and Bonferroni-corrected tests under the level 3 and level 4 sample size (10,000, 100,000) and significance level (0.05) criteria. It is clear that the relative effectiveness of the approaches is sensitive to the combinations of various parameters and scenario configurations.

Of the three methods only the Bonferroni succeeds in controlling the FWER at an acceptable level (recall that a device fails to meet the standard if a single point of leakage is discovered). Under the level 3 criteria it has lower power than the repetition in all leakage scenarios; however, at the level 4 sample size it is *more* powerful in the case that the density of leak points is fixed. In these fixed density cases the power of all the methods grows as the trace length increases; in the case that the *number* is fixed the unadjusted and repeated tests have a fixed overall probability of detection whilst the Bonferroni tests peak when there are no non-leaky points and then decrease at a speed which depends on the sample size. Note that, at level 4, the power to detect at least one of 30 leaks is still very close to 1 for traces of length up to 10 million; at level 3 it is close to zero from traces of 1 million or more.

At the TVLA significance level (see Fig. 7 in App. B) the FWER is (as we've already seen) still very low for both adjustment methods, even up to traces of length 10 million or more (not shown on the graph). The level 3 sample size is completely inadequate to detect effects of this size regardless of trace length. Interestingly, for the level 4 sample size the advantage displayed by the Bonferroni method has widened. We again see a decrease in power to detect a fixed number of leaks as the total length increases, however it should be pointed out that the power to detect one of at least 30 leaks is still above 0.999 for a trace of length 10 million (although it is lower than the power of the repetition step by this point).

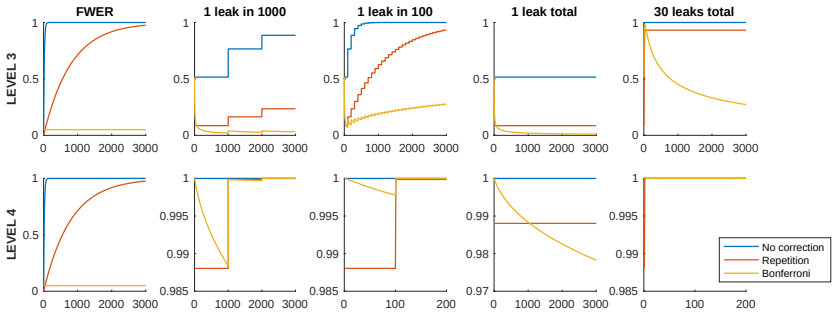


**Fig. 4.** FWER and 1-minimal ('overall') power of the tests to detect effects of the 'observed' size 0.04 for various leakage scenarios as the trace length increases, under the level 3 and level 4 standard criteria with a significance level of $\alpha = 0.05$. Note that some of the axes have been truncated in order to focus on the interesting regions of the graphs.

We remark that the level 4 standard criteria swapping the repetition step for the Bonferroni method seems an adequate measure to certify vulnerability and/or security for effect sizes of 0.04, even as the trace length increases. Swapping the significance level for the original TVLA recommendation of 0.00001 also

achieves this, although we note that the Bonferroni adjustment is anyway more powerful than the repetition step in this instance. However, we already know from Tab. 4 that the level 4 sample size is too small to reliably detect 'tiny' effects (repeating the Fig. 4 analysis confirms this and reveals no new insights). A reasonable question to ask is then what methods/parameter choices would enable certification with respect to these types of (still realistic) vulnerabilities.

As should be clear by now, appropriate configuration necessarily depends on the type of leakage scenario that we envisage. For example, a typical software implementation might produce very long (e.g. 100,000-point) traces; in the case that it is unprotected (and especially for the non-specific fixed-versus-random tests) the number of leak points could be high, say, 1 in 100; in the presence of countermeasures and/or in the case of a specific test the number could be far lower, say, 10 total, or even just one (which it remains crucial to be able to find). By contrast, hardware implementations are faster and typically produce shorter (e.g. 1,000-point) traces, with any leakage concentrated at one or a few indices.

Tab. 5 shows suitable parameter choices for Bonferroni-adjusted tests in each of these settings. The large sample sizes (especially when we are concerned with finding very sparse leakage) are something of a reality check on the popular view that leakage detection is a 'more efficient' alternative to performing attacks: the advantages of the former are best understood in terms of its potential to find a *wider variety* of possible sensitive dependencies than an attack-based approach. Meanwhile, precisely because an adversary is targeting a specific vulnerability – with a tailored tool, using information (if available) about the form of the data dependency – we should always expect attacks to be more data efficient than detection tests. It follows (importantly) that we should never interpret the sample sizes required for leakage detection as quantitative markers of a device's resistance to attack. Reciprocally, attack-based configurations should not be used to inform the specifications of detection-based approaches: the influence of the (originally attack-based) FIPS 140-2 on the (detection-based) ISO 17825 likely explains why the level 3 and 4 sample sizes are as limitingly small as they are.

| **Scenario type** | **Trace** | **#** | **ISO** $\alpha = 0.05$ | | **TVLA** $\alpha = 0.00001$ | |
|---|---|---|---|---|---|---|
| | **length** | **leaks** | $d = 0.04$ | $d = 0.001$ | $d = 0.04$ | $d = 0.001$ |
| Software (generic leaks) | 100,000 | 100 | $2.5 \times 10^4$ | $3.9 \times 10^7$ | $6.8 \times 10^4$ | $1.1 \times 10^8$ |
| Software (specific leaks) | 100,000 | 10 | $4.8 \times 10^4$ | $7.7 \times 10^7$ | $1.2 \times 10^5$ | $1.9 \times 10^8$ |
| Software (protected) | 100,000 | 1 | $1.1 \times 10^5$ | $1.8 \times 10^8$ | $2.9 \times 10^5$ | $4.6 \times 10^8$ |
| Hardware (unprotected) | 1,000 | 10 | $2.9 \times 10^4$ | $4.6 \times 10^7$ | $9.6 \times 10^4$ | $1.5 \times 10^8$ |
| Hardware (protected) | 1,000 | 1 | $8.1 \times 10^4$ | $1.3 \times 10^8$ | $2.5 \times 10^5$ | $4.0 \times 10^8$ |

**Table 5.** Parameter combinations for reliably certifying vulnerability/security in different realistic leakage scenarios using the Bonferroni adjustment to control the false positive rate at an overall level $\alpha$.

**Remark:** At this point it is important to recall that in an actual evaluation the entire process has to be applied to several/many intermediate values as part of the specific detection tests. These further tests are synonymous with considering longer traces and an extended analysis would be possible given a specified number of tests.

## 7 Conclusions and Recommendations

TVLA was originally conceived as a structured set of leakage detection tests to overcome the issue of having to test against an ever increasing number of attack vectors (thus the concern was coverage of an evaluation rather than trace efficiency). An in-dept statistical analysis was never carried out yet these recommendations became the basis for leakage evaluations as specified in ISO 17825.

We have shown that **following the ISO 17825 recommendations to the letter would result in the failure of *all* target devices** (at security levels 3 and 4) with extremely high probability. This is because of the inflation of Type I errors (false positives) as the number of jointly performed statistical hypothesis tests increases.

The problem can be mitigated by **replacing the (somewhat ad hoc) test repetition step (inherited from TVLA) with an established statistical method** to control the overall error rate, such as the Bonferroni adjustment, and/or by replacing the threshold for significance with the stricter one originally implied by the TVLA standard. In the latter case, the repetition step is anyway shown to be less efficient than Bonferroni-style adjustments, so we recommend against adhering to that part of TVLA.

There are some ambiguities in ISO 17825 about how to interpret the acquisition criterion. Even opting for the most generous interpretation, the **level 3 sample size specification is shown to be inadequate to certify vulnerability/security against effects of the size and frequency that we observe in a range of typical 'easy to attack' implementations**. The level 4 specification *is* able to detect these with high probability, even with the stricter TVLA-based significance threshold provided the leakages are of sufficient density as the length of the trace increases. **However, neither are sufficient to detect the types of 'tiny' effects that have been shown to exist (and to be exploitable) by larger-scale academic studies**.

We therefore recommend the **necessity for larger acquisitions** than those specified by the standard. A difficulty here is that, although statistical power analysis provides tools to derive the appropriate sample sizes for a particular test scenario, it requires considerable *a priori* information about that scenario to do so (even more so in the case of multiple tests and their corresponding adjustment procedures). Whilst it is possible to broadly identify common expected features across classes of scenario, **a preferable approach would be to develop a two-stage evaluation procedure** combining an exploratory phase with a pared-down confirmatory analysis in which information about the covariance structure and likely location/nature of the leaks is used to inform

the acquisition process and to chose a (reduced set) of carefully-formulated hypothesis tests to perform. We leave the precise details of such a strategy as an interesting avenue for further work.

However the standard procedures (or adaptations therefore) are applied it is **important that outcomes are presented responsibly**. An evaluator needs to decide – and to give a justification for – the false positive and false negative rates that are acceptable. For example, even if a multiplicity adjustment is used to successfully control the overall false positive rate at the level specified by the standard, this still implies that 5 in every 100 secure devices will fail the test at random. If this is considered too high, then a stricter significance criterion will need to be chosen, inevitably implying greater data complexity. Either way, **the error rates must be made transparent – as should the effect size** the test is able to detect, **the coverage limitations** that we identified in Section 5.1, and the fact that the sample size needed for a successful *attack* may be much smaller than that required for detection.

*Acknowledgements* Our work has been funded by the European Commission through the H2020 project 731591 (acronym REASSURE). A fuller report on this aspect of the project can be found in *A Cautionary Note Regarding the Usage of Leakage Detection Tests in Security Evaluation* [42].

## References

1. P. Ammann and J. Offutt. *Introduction to Software Testing*. Cambridge University Press, New York, NY, USA, 1 edition, 2008.
2. D. Asonov and R. Agrawal. Keyboard acoustic emanations. In *IEEE Symposium on Security and Privacy*, pages 3–11. IEEE Computer Society, 2004.
3. S. Bhasin, J. Danger, S. Guilley, and Z. Najm. Side-channel leakage and trace compression using normalized inter-class variance. In R. B. Lee and W. Shi, editors, *HASP 2014, Hardware and Architectural Support for Security and Privacy*, pages 7:1–7:9. ACM, 2014.
4. R. Bi and P. Liu. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics*, 17(1):146, Mar 2016.
5. J. Brouchier, T. Kean, C. Marsh, and D. Naccache. Temperature Attacks. *IEEE Security & Privacy*, 7(2):79–82, 2009.
6. K. Chatzikokolakis, T. Chothia, and A. Guha. Statistical Measurement of Information Leakage. In *TACAS*, pages 390–404, 2010.
7. T. Chothia and A. Guha. A Statistical Test for Information Leaks Using Continuous Mutual Information. In *CSF*, pages 177–190, 2011.
8. J. Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 1988.
9. J.-L. Danger, G. Duc, S. Guilley, and L. Sauvage. Education and open benchmarking on side-channel analysis with the DPA contests. In *NIST Non-invasive attack testing workshop*, 2011.
10. T. De Cnudde, M. Ender, and A. Moradi. Hardware Masking, Revisited. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2018(2):123–148, May 2018.

11. A. A. Ding, L. Zhang, F. Durvaux, F.-X. Standaert, and Y. Fei. Towards sound and optimal leakage detection procedure. In T. Eisenbarth and Y. Teglia, editors, *Smart Card Research and Advanced Applications*, pages 105–122. Springer International Publishing, 2018.
12. O. J. Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.
13. F. Durvaux and F. Standaert. From Improved Leakage Detection to the Detection of Points of Interests in Leakage Traces. In M. Fischlin and J. Coron, editors, *Advances in Cryptology – EUROCRYPT 2016*, volume 9665 of *LNCS*, pages 240–262. Springer, 2016.
14. B. Efron. Size, power and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377, 08 2007.
15. J. Ferrigno and M. Hlaváč. When AES blinks: Introducing optical side channel. *IET Information Security*, 2(3):94–98, 2008.
16. K. Gandolfi, C. Mourtel, and F. Olivier. Electromagnetic Analysis: Concrete Results. In Ç. K. Koç, D. Naccache, and C. Paar, editors, *Proceedings of CHES 2001*, volume 2162 of *LNCS*, pages 251–261. Springer, 2001.
17. G. Goodwill, B. Jun, J. Jaffe, and P. Rohatgi. A testing methodology for side-channel resistance validation. In *NIST Non-invasive attack testing workshop*, 2011.
18. J. M. Hoenig and D. M. Heisey. The Abuse of Power. *The American Statistician*, 55(1):19–24, 2001.
19. S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 01 1979.
20. Information technology – Security techniques – Testing methods for the mitigation of non-invasive attack classes against cryptographic modules. Standard, International Organization for Standardization, Geneva, CH, 2016.
21. Information technology – Security techniques – Security requirements for cryptographic modules. Standard, International Organization for Standardization, Geneva, CH, 2012.
22. P. C. Kocher. Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems. In N. Koblitz, editor, *Advances in Cryptology – CRYPTO '96*, volume 1109 of *LNCS*, pages 104–113. Springer, 1996.
23. P. C. Kocher, J. Jaffe, and B. Jun. Differential Power Analysis. In *CRYPTO*, pages 388–397, 1999.
24. P. Liu and J. T. G. Hwang. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23(6):739–746, 2007.
25. L. Mather, E. Oswald, J. Bandenburg, and M. Wójcik. Does My Device Leak Information? An a priori Statistical Power Analysis of Leakage Detection Tests. In K. Sako and P. Sarkar, editors, *Advances in Cryptology – ASIACRYPT 2013*, volume 8269 of *LNCS*, pages 486–505. Springer, 2013.
26. J. C. Miller and C. J. Maloney. Systematic Mistake Analysis of Digital Computer Programs. *Commun. ACM*, 6(2):58–63, Feb. 1963.
27. K. E. Porter. Statistical power in evaluations that investigate effects on multiple outcomes: A guide for researchers. *Journal of Research on Educational Effectiveness*, 0(0):1–29, 2017.
28. S. Pounds and C. Cheng. Sample size determination for the false discovery rate. *Bioinformatics*, 21(23):4263–4271, 2005.
29. J.-J. Quisquater and D. Samyde. ElectroMagnetic Analysis (EMA): Measures and Counter-measures for Smart Cards. In I. Attali and T. Jensen, editors, *Smart Card*

*Programming and Security*, volume 2140 of *LNCS*, pages 200–210. Springer Berlin / Heidelberg, 2001.
30. S. S. Sawilowsky. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599, 2009.
31. W. Schindler, K. Lemke, and C. Paar. A Stochastic Model for Differential Side Channel Cryptanalysis. In J. R. Rao and B. Sunar, editors, *Proceedings of CHES 2005*, volume 3659 of *LNCS*, pages 30–46. Springer, 2005.
32. T. Schneider and A. Moradi. Leakage Assessment Methodology – A Clear Roadmap for Side-Channel Evaluations. In T. Güneysu and H. Handschuh, editors, *Proceedings of CHES 2015*, volume 9293 of *LNCS*, pages 495–513. Springer, 2015.
33. A. Shamir and E. Tromer. Acoustic cryptanalysis (website). `http://theory.csail.mit.edu/~tromer/acoustic/`. (Accessed 9th September 2019).
34. S. Skorobogatov. Using Optical Emission Analysis for Estimating Contribution to Power Analysis. In L. Breveglieri, I. Koren, D. Naccache, E. Oswald, and J.-P. Seifert, editors, *Fault Diagnosis and Tolerance in Cryptography – FDTC '09*, pages 111–119. IEEE Computer Society, 2009.
35. F.-X. Standaert, B. Gierlichs, and I. Verbauwhede. Partition vs. Comparison Side-Channel Distinguishers: An Empirical Evaluation of Statistical Tests for Univariate Side-Channel Attacks against Two Unprotected CMOS Devices. In *ICISC*, pages 253–267, 2008.
36. F.-X. Standaert, O. Pereira, Y. Yu, J.-J. Quisquater, M. Yung, and E. Oswald. Leakage Resilient Cryptography in Practice. In A.-R. Sadeghi and D. Naccache, editors, *Towards Hardware-Intrinsic Security: Foundations and Practice*, pages 99–134. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
37. S. Tasiran and K. Keutzer. Coverage metrics for functional validation of hardware designs. *IEEE Des. Test*, 18(4):36–45, July 2001.
38. A. Thillard, E. Prouff, and T. Roche. Success through Confidence: Evaluating the Effectiveness of a Side-Channel Attack. In G. Bertoni and J.-S. Coron, editors, *Cryptographic Hardware and Embedded Systems – CHES 2013*, pages 21–36, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
39. T. Tong and H. Zhao. Practical guidelines for assessing power and false discovery rate for fixed sample size in microarray experiments. *Statistics in medicine*, 27:1960–72, 05 2008.
40. Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.
41. B. L. Welch. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947.
42. C. Whitnall and E. Oswald. A Cautionary Note Regarding the Usage of Leakage Detection Tests in Security Evaluation. IACR Cryptology ePrint Archive, Report 2019/703, 2019. `https://eprint.iacr.org/2019/703`.

## A Sample Size for the *t*-Test

We begin with a simple visual example that illustrates the concepts of $\alpha$ and $\beta$ values and their relationship to the sample size.

Consider the following two-sided hypothesis test for the mean of a Gaussian-distributed variable $A \sim \mathcal{N}(\mu, \sigma)$, where $\mu$ and $\sigma$ are the (unknown) parameters:

$$H_0 : \mu = \mu_0 \text{ vs. } H_{alt} : \mu \neq \mu_0. \tag{3}$$

Note that, in the leakage detection setting, where one typically wishes to test for a non-zero difference in means between *two* Gaussian distributions $Y_1$ and $Y_2$, this can be achieved by defining $A = Y_1 - Y_2$ and (via the properties of the Gaussian distribution) performing the above test with $\mu_0 = 0$.

Suppose the alternative hypothesis is true and that $\mu = \mu_{alt}$. This is called a 'specific alternative'[10], in recognition of the fact that it is not usually possible to compute power for *all* the alternatives when $H_{alt}$ defines a set or range. In the leakage detection setting one typically chooses $\mu_{alt} > 0$ to be the smallest difference $|\mu_1 - \mu_2|$ that is considered of practical relevance; this is called the effect size. Without loss of generality, we suppose that $\mu_{alt} > \mu_0$.

Figure 5 illustrates the test procedure when the risk of a Type I error is set to $\alpha$ and the sample size is presumed large enough (typically $n > 30$) that the distributions of the test statistic under the null and alternative hypotheses can be approximated by Gaussian distributions. The red areas together sum to $\alpha$; the blue area indicates the overlap of $H_0$ and $H_{alt}$ and corresponds to $\beta$ (the risk of a Type II error). The power of the test – that is, the probability of correctly rejecting the null hypothesis when the alternative in true – is then $1 - \beta$, as depicted by the shaded area.

There are essentially three ways to raise the power of the test. One is to increase the effect size of interest which, as should be clear from Figure 5, serves to push the distributions apart, thereby diminishing the overlap between them. Another is to increase $\alpha$ – that is, to make a trade-off between Type II and Type I errors – or (if appropriate) to perform a one-sided test, either of which has the effect (in this case) of shifting the critical value to the left so that the shaded region becomes larger. (In the leakage detection case the one-sided test is unlikely to be suitable as differences in either direction are equally important and neither can be ruled out *a priori*). The third way to increase the power is to increase the sample size for the experiment. This reduces the standard error on the sample means, which again pushes the alternative distribution of the test statistic further away from null (note from Figure 5 that it features in the denominator of the distance).

Suppose you have an effect size in mind – based either on observations made during similar previous experiments, or on a subjective value judgement about how large an effect needs to be before it is practically relevant (e.g. the level of leakage which is deemed intolerable) – and you want your test to have a given confidence level $\alpha$ and power $1 - \beta$. The relationship between confidence, power, effect size and sample size can then be used to derive the minimum sample size necessary to achieve this.

The details of the argumentation that now follows are specific to a two-tailed $t$-test, but the general procedure can be adapted to any test for which the distribution of the test statistic is known under the null and alternative hypotheses.

[10] The overloading of terminology between 'specific alternatives' and 'specific' TVLA tests is unfortunate but unavoidable.

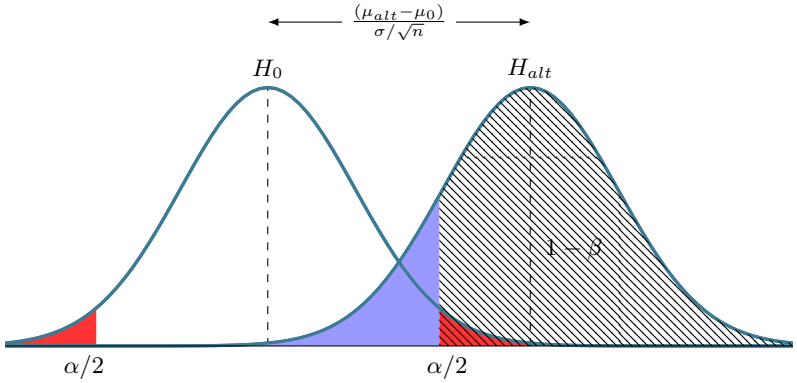


**Fig. 5.** Figure showing the Type I and II error probabilities, $\alpha$ and $\beta$ as well as the effect size $\mu_{alt} - \mu_0$ for a specific alternative such that $\mu_{alt} > \mu_0$.

For the sake of simplicity (i.e. to avoid calculating effectively irrelevant degrees of freedom) we will assume that our test will in any case require the acquisition of more than 30 observations, so that the Gaussian approximations for the test statistics hold as in Figure 5. Without loss of generality we also assume that the difference of means is positive (otherwise the sets can be easily swapped). Finally, we assume that we seek to populate both sets with equal numbers $n = |Y|/2$ of observed traces.

**Theorem 1.** *Let $Y_1$ be a set of traces of size $N/2$ drawn via repeat sampling from a normal distribution $\mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2$ be a set of traces of size $N/2$ drawn via repeat sampling from a normal distribution $\mathcal{N}(\mu_2, \sigma_2^2)$. Then, in a two-tailed test for a difference between the sample means:*

$$H_0\text{: } \mu_1 = \mu_2 \text{ vs. } H_{alt}\text{: } \mu_1 \neq \mu_2, \tag{4}$$

*in order to achieve significance level $\alpha$ and power $1 - \beta$, the overall number of traces $N$ needs to be chosen such that:*

$$N \geq 2 \cdot \frac{(z_{\alpha/2} + z_\beta)^2 \cdot ({\sigma_1}^2 + {\sigma_2}^2)}{(\mu_1 - \mu_2)^2}. \tag{5}$$

Note that Equation 5 can be straightforwardly rearranged to alternatively compute any of the significance level, effect size or power in terms of the other three quantities.

## B Results for Original TVLA-Recommended Threshold

| Cohen's $d$ | Power | | Power | Cohen's $d$ | |
|---|---|---|---|---|---|
| | Level 3 | Level 4 | | $N = 10,000$ | $N = 100,000$ |
| Very small (0.01) | 0.000 | 0.002 | 0.75 | 0.102 | 0.032 |
| Small (0.2) | 1.000 | 1.000 | 0.80 | 0.105 | 0.033 |
| Medium (0.5) | 1.000 | 1.000 | 0.90 | 0.114 | 0.036 |
| Large (0.8) | 1.000 | 1.000 | 0.95 | 0.121 | 0.038 |
| Very large (1.2) | 1.000 | 1.000 | 0.99 | 0.135 | 0.043 |
| Huge (2) | 1.000 | 1.000 | 0.99999 | 0.174 | 0.055 |

**Table 6.** LHS: Power to achieve Cohen's and Sawilowsky's standardised effects under the TVLA significance criteria (which approximates to $\alpha = 0.00001$) and the standard level 3 ($N = 10,000$) and level 4 ($N = 100,000$) sample size criteria; RHS: Minimum effect sizes detectable for increasing power thresholds.

| Effect | Repeat test? | Level 3 | | Level 4 | | Required sample size | |
|---|---|---|---|---|---|---|---|
| | | Ave | 1-min | Ave | 1-min | Ave | 1-min |
| 0.04 | No | 0.008 | 0.210 | 0.972 | 1.000 | 188,446 | 38,924 |
| 0.04 | Yes | 0.000 | 0.000 | 0.272 | 1.000 | 390,228 | 104,867 |
| 0.001 | No | 0.000 | 0.000 | 0.000 | 0.000 | 301,512,956 | 62,279,197 |
| 0.001 | Yes | 0.000 | 0.000 | 0.000 | 0.000 | 624,365,394 | 167,786,951 |

**Table 7.** Average ('per-test') and 1-minimal ('overall') power to detect observed and 'tiny' effect sizes under the level 3 and 4 criteria, and the sample size required to achieve balanced errors for a significance criterion of $\alpha = 0.00001$. (30 leak points in a trace set of length 1,400).

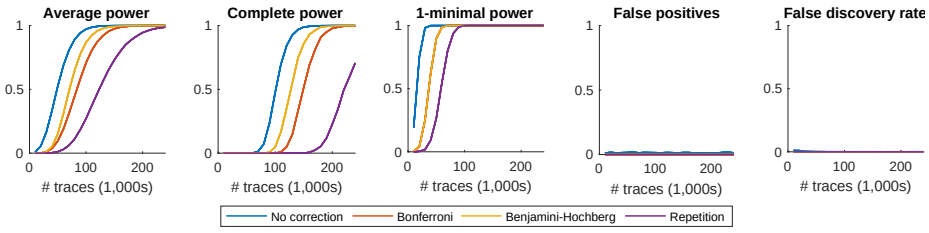


**Fig. 6.** Different types of power and error to detect 30 true effects of size 0.04 in a trace set of length 1,400, as sample size increases, for an overall significance level of $\alpha = 0.00001$. (Based on 5,000 random draws from the multivariate test statistic distribution under the alternative hypothesis).

| Correction strategy | Level 3 | | Level 4 | |
|---|---|---|---|---|
| | **1-min power** | **FWER** | **1-min power** | **FWER** |
| None | 0.1912 | 0.0156 | 1.0000 | 0.0134 |
| Bonferroni | 0.0020 | 0.0000 | 1.0000 | 0.0000 |
| Šidák | 0.0020 | 0.0000 | 1.0000 | 0.0000 |
| Holm | 0.0020 | 0.0000 | 1.0000 | 0.0000 |
| Benjamini-Hochberg | 0.0020 | 0.0000 | 1.0000 | 0.0000 |
| Repetition | 0.0000 | 0.0000 | 0.9986 | 0.0000 |

**Table 8.** Different types of power and error to detect 30 true effects of size 0.04 in a trace set of length 1,400, under the level 3 and level 4 sample size criteria and with an overall significance level of $\alpha = 0.00001$. (Based on 5,000 random draws from the multivariate test statistic distribution under the alternative hypothesis).

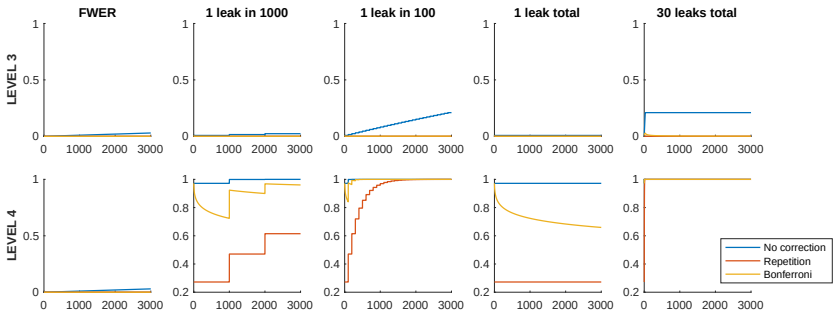


**Fig. 7.** FWER and 1-minimal ('overall') power of the tests to detect effects of the 'observed' size 0.04 for various leakage scenarios as the trace length increases, under the level 3 and level 4 standard criteria with an overall significance level of $\alpha = 0.00001$.