# The effectiveness of sampling plans based on both item-by-item tests and the stratum D-statistic

**Thomas Krieger**[1]
Forschungszentrum Jülich GmbH, Jülich, Germany

**Aaron M. Bevill, Robert Binner, Sarah Michalak and Claude F. Norman**
International Atomic Energy Agency, Vienna, Austria

***Abstract***

*Nuclear fuel cycle facility declarations on nuclear material inventories and transfers are independently verified by the IAEA. These verification activities usually rely on a sampling plan that is designed to achieve a specified probability to detect falsification of operator reports. Currently, the IAEA's sampling plans assume item-by-item tests in which the difference between the reported and the measured value of each item selected for verification is compared to a threshold. If a difference exceeds this threshold, then an "alarm" occurs, and the cause for the difference is further investigated.*

*In the present paper we analyse sampling plans in which in addition to the usual item-by-item tests, a stratum difference statistic of the verified items is applied as a test statistic. The reason for considering the stratum difference statistic in addition to the item-by-item tests is that it is "better" at detecting bias defect falsifications than the item-by-item tests. Therefore, we investigate the effectiveness in terms of the achieved detection probability of sampling plans in which both tests are applied and analyse whether sample sizes could be reduced while still achieving the required detection probability.*

**Keywords:** IAEA; Verification sampling plans; item-by-item tests; difference statistic

## 1 Introduction and Motivation

Nuclear fuel cycle facility declarations on nuclear material inventories and transfers are independently verified by the IAEA. These verification activities usually rely on a sampling plan that is designed to achieve a specified probability to detect falsification of operator reports by a specified total amount of material. Currently, the IAEA's sampling plans assume item-by-item tests (shortly $SP_{item}$) that are based on item-by-item tests in which the relative difference between the reported and the measured value of each item selected for verification is compared to a threshold. If a relative difference exceeds this threshold, then an "alarm" occurs, and the cause for the difference is further investigated. This clarification is incorporated into the probabilistic model underlying $SP_{item}$. Under further assumptions (see [1] and section 2), the detection probability (DP) can be derived, and a sampling plan can be determined that achieves a required DP. This approach is summarized in section 2.

Instead of or in addition to the item-by-item tests, the overall (over all observed relative differences) relative difference statistic ($D$-statistic) of the verified items could be applied as test statistic. These sampling plans are abbreviated by $SP_D$ and $SP_{item,D}$, and they are investigated in sections 3 and 4. It will turn out that if the current IAEA detection event of "observing at least one significant (item-by-item) difference for a falsified item" is generalized to the combined event of "observing at least one significant (item-by-item) difference for a falsified item" *or* "observing a significant relative stratum D-

[1] Corresponding author. Institute of Energy and Climate Research, IEK-6: Nuclear Waste Management and Reactor Safety, 52425 Jülich, Germany, t.krieger@fz-juelich.de

statistic when there is at least one falsified item among the sampled items", then $SP_{item,D}$ leads – at least for the numerical examples considered so far – to a remarkable reduction of the sample size while still achieving the required detection probability.

Section 5 concludes the paper with a summary and future research activities.

## 2 Sampling plans based on item-by-item tests ($SP_{item}$)

Let

- $N$ be the number of items in the stratum for which the operator has declared a mass;
- $\bar{x}$ be the average amount of declared nuclear material in an item as estimated from the operator's declaration;
- $M$ be the diverter's goal quantity (usually, one significant quantity);
- $\delta$ be the combined relative standard deviation (RSD) of the operator's and inspector's verification measurement methods;
- $n$ be the sample size, i.e., the number of items to be verified by the inspector;
- $r$ be the number of falsified items in the stratum.

Current IAEA sampling plans are based on several assumptions, which are described in detail in [1]. To name just a few of these assumptions: Purely random multiplicative measurement model, diversion into the relative difference statistic (see below), item-by-item testing, the goal quantity $M$ is diverted from the stratum, and equal diversion hypothesis, i.e., if $r$ items are falsified, then the amount $M/r$ is or will be removed from each of these items.

Because it is assumed that all falsified items are falsified by the same amounts (see the equal diversion hypothesis above), the set of diversion strategies is given by $\{\lceil M/\bar{x} \rceil, \dots, N\}$, where the ceiling operator $\lceil a \rceil$ of a real number $a$ is the smallest integer not less than $a$.

In the accounting records, the operator declares the nuclear material masses $x_1, \dots, x_N$, where $x_i$ is the mass of the $i$-th item, and $\bar{x}$ the average item mass, i.e., the population is assumed to be sufficiently homogeneous; see [2].

Because a purely random multiplicative measurement model is assumed, these masses are modelled as realizations of a random variables $X_i = t_{Op,i}(1 + R_{Op,i})$, where $t_{Op,i}$ is the true, but unknown, amount of nuclear material of item $i$ and $R_{Op,i}$ is the normally distributed random error with expectation zero and variance $\delta_{Op}^2$, i.e., $R_{Op,i} \sim \mathcal{N}(0, \delta_{Op}^2)$ for $i = 1, \dots, N$.

The inspector verifies $n$ out of the $N$ items by performing independent measurements. Without loss of generality (see [3] or [4]) it can be assumed that the first $n$ items are verified, i.e., the inspector measures $Y_i = t_{Insp,i}(1 + R_{In,i})$, where $t_{Insp,i}$ is the true, but unknown, amount of nuclear material present in item $i$ and $R_{In,i} \sim \mathcal{N}(0, \delta_{In}^2)$ for $i = 1, \dots, n$. Let $\delta^2 := \delta_{Op}^2 + \delta_{In}^2$.

Using the individual relative differences $D_i := (X_i - Y_i)/X_i$ for $i = 1, \dots, n$, the inspector performs a statistical test on the item-by-item basis to test the hypothesis

$$H_0: \text{no diversion} \quad \text{versus} \quad H_1: \text{diversion of the goal quantity } (M) \text{ from the stratum},$$

i.e., he compares the individual $D_i$ to a threshold. If under $H_1$ $i$ falsified items are in the sample of size $n$, then it can also be assumed without loss of generality (see [3] or [4]) that the first $i$ items are the falsified ones, because the equal diversion hypothesis implies that the DP depends on the number of falsified items in the sample, and not exactly which ones are falsified.

Thus, we have for the overstatement case (i.e., less material is present than is declared, see [1]) and under some suitable assumptions (see [1])

$$D_j \sim \mathcal{N}(0,\delta^2) \quad \text{for} \quad j=1,\dots,n \qquad \text{under} \quad H_0$$
$$D_j \sim \begin{cases} \mathcal{N}\left(\frac{M}{\bar{x}r},\left(1-\frac{M}{\bar{x}r}\right)^2\delta^2\right) & \text{for} \quad j=1,\dots,i \\ \mathcal{N}(0,\delta^2) & \text{for} \quad j=i+1,\dots,n \end{cases} \qquad \text{under} \quad H_1 \,. \tag{1}$$

Eq. (1) indicates, that even under $H_1$ some of the items (namely $D_{i+1},\dots,D_n$) are distributed as if $H_0$ were true.

Considering the threshold $3\delta$, a false alarm (FA) is raised if $D_j > 3\delta$ for at least one $j=1,\dots,n$. This decision rule leads to a single FAP of approximately $.0013$ because (assume one-sided testing)

$$\mathbb{P}_{H_0}\left(D_j > 3\delta\right) = 1-\mathbb{P}_{H_0}\left(\frac{D_j}{\delta}\le 3\right) = 1-\Phi(3) \approx .0013\,,$$

where $\Phi(\cdot)$ is the distribution function of the standard normally distributed random variable. Because a purely random multiplicative measurement model is assumed, the false alarm probability FAP $\alpha_1(n)$ is

$$\begin{aligned} \alpha_1(n) &:= \mathbb{P}_{H_0}(\text{at least one false alarm}) = 1-\mathbb{P}_{H_0}(\text{no false alarms}) \\ &= 1-\mathbb{P}_{H_0}(D_1\le 3\delta,\dots,D_n\le 3\delta) = 1-\left(\Phi(3)\right)^n . \end{aligned} \tag{2}$$

Under $H_1$ (diversion of the goal quantity $M$ from the stratum), we have $i$ falsified and $n-i$ non-falsified items in the sample ($1\le i\le Min(r,n)$), i.e., by Eq. (1): $D_1,\dots,D_i \sim \mathcal{N}(M/(\bar{x}r),(1-M/(\bar{x}r))^2\delta^2)$ and $D_{i+1},\dots,D_n \sim \mathcal{N}(0,\delta^2)$. In [1] it is discussed in detail that any alarm, i.e., any significant individual relative difference, is clarified as to whether it is due to a false alarm or due to a diversion. Thus, only the differences $D_1,\dots,D_i$ of the falsified items are considered, while the differences $D_{i+1},\dots,D_n$ of the non-falsified items – which may only lead to a false alarm – are excluded.

Therefore, the non-identification probability, i.e., the probability of not identifying/classifying any falsified item as falsified, is given by

$$\begin{aligned} \mathbb{P}_{H_1}(D_1\le 3\delta,\dots,D_i\le 3\delta) &= \prod_{j=1}^{i}\mathbb{P}_{H_1}\left(\frac{D_j-M/(\bar{x}r)}{(1-M/(\bar{x}r))\delta}\le\frac{3\delta-M/(\bar{x}r)}{(1-M/(\bar{x}r))\delta}\right) \\ &= \left(\Phi\left(\frac{3\delta-M/(\bar{x}r)}{(1-M/(\bar{x}r))\delta}\right)\right)^i . \end{aligned} \tag{3}$$

Because the IAEA verification sampling plans are based on item-by-item tests, the detection event is defined by

$$\text{detection event} := \{\text{observing at least one significant item}-\text{by}-\text{item relative difference for a falsified item}\} \,. \tag{4}$$

Without going into the details (see [1]), using the hypergeometric distribution, Eqs. (3) and (4) and the law of total probability (see [5] or [6]) yield for the detection probability $DP(n,r)$

$$\begin{aligned} DP(n,r) &:= \mathbb{P}_{H_1}(\text{detection event}) \\ &= 1-\sum_{i=Max\left(0,r-(N-n)\right)}^{Min(r,n)}\frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}}\left(\Phi\left(\frac{3\delta-M/(\bar{x}r)}{(1-M/(\bar{x}r))\delta}\right)\right)^i . \end{aligned} \tag{5}$$

For example, let

$$N=250, \quad \bar{x}=5, \quad \delta=1\%, \quad M=75\ [\text{kg}]\,. \tag{6}$$

The DP curve is depicted in Figure 1 for the sample size $n=27$, that is the smallest sample size such that $DP(n,r)\ge 0.2$ for all $r$ from the set $\{\lceil M/\bar{x}\rceil,\dots,N\}$ of diversion strategies.

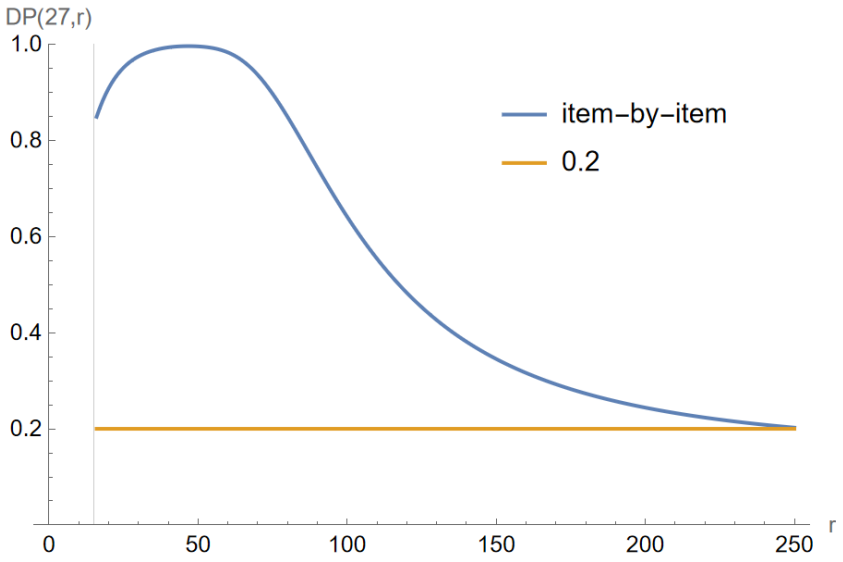


Figure 1: $DP(27, r)$ as a function of the number $r$ of falsified items.

The left-hand vertical line in Figure 1 refers to gross defect falsification in which $\lceil M/\bar{x} \rceil = 15$ items are assumed to be completely emptied.

## 3 Sampling plans based on the relative stratum difference $D$-statistic ($SP_D$)

The sampling plans considered in this section are based on the relative stratum difference statistic defined by

$$D := \frac{N}{n} \sum_{j=1}^{n} D_j \,, \tag{7}$$

i.e., instead of comparing the $n$ differences $D_j$ individually as in section 2, they are summed up and propagated to the whole stratum. Therefore, the inspector's decision is based on aggregated information.

If $H_0$ (no diversion) is true, then (recall: purely random multiplicative measurement model) Eq. (7) yields, using Eq. (1),

$$D \sim \mathcal{N}\left(0, \frac{N^2}{n}\delta^2\right). \tag{8}$$

Let $U(\cdot)$ denotes the inverse function of $\Phi(\cdot)$. The threshold $k_D$ with which the value of the $D$-statistic is compared is given by

$$k_D = \frac{N}{\sqrt{n}}\, \delta\, U\Big(\big(\Phi(3)\big)^n\Big)\,, \tag{9}$$

because then $SP_{item}$ and $SP_D$ have the same FAP: By Eqs. (2) and Eq. (8) we have

$$\begin{aligned} \alpha_2(n) \;&:=\; \mathbb{P}_{H_0}\left(D > \frac{N}{\sqrt{n}}\, \delta\, U\Big(\big(\Phi(3)\big)^n\Big)\right) = 1 - \mathbb{P}_{H_0}\left(\frac{\sqrt{n}\, D}{N\, \delta} \le U\Big(\big(\Phi(3)\big)^n\Big)\right) \\ &=\; 1 - \big(\Phi(3)\big)^n = \alpha_1(n)\,. \end{aligned} \tag{10}$$

Suppose $H_1$ is true (diversion of the goal quantity $M$ from the stratum). If there are $1 \le i \le Min(r, n)$ falsified items in the sample, then the expectation and variance of the relative stratum $D$-statistic are by Eq. (1)

$$\mathbb{E}_{H_1}(D) = \mathbb{E}_{H_1}\left(\frac{N}{n}\sum_{j=1}^{n} D_j\right) = \frac{N}{n}\; i\; \frac{M}{\bar{x}r} \qquad \text{and}$$

$$\mathbb{V}_{H_1}(D) = \frac{N^2}{n^2}\sum_{j=1}^{n} \mathbb{V}_{H_1}(D_j) = \frac{N^2}{n^2}\left(i\left(1-\frac{M}{\bar{x}r}\right)^2 \delta^2 + (n-i)\,\delta^2\right),$$

therefore, the probability of observing a non-significant relative stratum $D$-statistic is

$$\begin{aligned} & \mathbb{P}_{H_1}(D \le k_D) \\ = \;& \mathbb{P}_{H_1}\left(\frac{D - \frac{N}{n}\; i\; \frac{M}{\bar{x}r}}{\frac{N}{n}\sqrt{i\left(1-\frac{M}{\bar{x}r}\right)^2 \delta^2 + (n-i)\,\delta^2}} \le \frac{k_D - \frac{N}{n}\; i\; \frac{M}{\bar{x}r}}{\frac{N}{n}\sqrt{i\left(1-\frac{M}{\bar{x}r}\right)^2 \delta^2 + (n-i)\,\delta^2}}\right) \\ = \;& \Phi\left(\frac{\sqrt{n}\; U\left(\left(\Phi(3)\right)^n\right) - i\; \frac{M}{\bar{x}r\delta}}{\sqrt{i\left(1-\frac{M}{\bar{x}r}\right)^2 + n - i}}\right). \end{aligned} \tag{11}$$

In contrast to Eq. (3), Eq. (11) is not just based on the differences $D_1, \dots, D_i$ of the falsified item but also on the differences $D_{i+1}, \dots, D_n$ of the non-falsified items.

Because sampling plans based on the relative stratum $D$-statistic are not (yet) applied by the IAEA, there exist no agreed rules for clarifying an alarm raised by applying the relative stratum $D$-statistic, and we assume (as for the item-by-item tests in section 2) in the following that any significant value of the relative stratum $D$-statistic is clarified as to whether it is due to a false alarm or due to a diversion: In case of $d \le k_D$ (non-significant value of the relative stratum $D$-statistic) no further action is needed, and in case of $d > k_D$ the significant relative stratum $D$-statistic is attributed to a diversion. We also assume that the detection event is redefined as follows:

$$\begin{aligned} \text{significant stratum } D-\text{statistic} \;\coloneqq\; \{&\text{observing a significant value of the relative} \\ &\text{stratum } D-\text{statistic when there is at least} \\ &\text{one falsified item among the sampled items}\} \end{aligned} \quad . \tag{12}$$

The probability of the event in Eq. (12) is, in analogy to Eq. (5) and by using Eq. (11), given by

$$\begin{aligned} & P_D(n,r) = \mathbb{P}_{H_1}(\text{significant stratum } D-\text{statistic}) \\ \coloneqq\; & \begin{cases} 1 - \sum_{i=r-(N-n)}^{Min(r,n)} \frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}}\; \Phi\left(\frac{\sqrt{n}\; U\left(\left(\Phi(3)\right)^n\right) - i\; \frac{M}{\bar{x}r\delta}}{\sqrt{i\left(1-\frac{M}{\bar{x}r}\right)^2 + n - i}}\right) & \\ \qquad\qquad \text{for } r-(N-n) \ge 1 & \\ 1 - \frac{\binom{N-r}{n}}{\binom{N}{n}} - \sum_{i=1}^{Min(r,n)} \frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}}\; \Phi\left(\frac{\sqrt{n}\; U\left(\left(\Phi(3)\right)^n\right) - i\; \frac{M}{\bar{x}r\delta}}{\sqrt{i\left(1-\frac{M}{\bar{x}r}\right)^2 + n - i}}\right) & \\ \qquad\qquad \text{for } r-(N-n) \le 0\;. & \end{cases} \end{aligned} \tag{13}$$

Note that $P_D(n,r)$ is not called detection probability, because the underlying event in Eq. (12) is different from that of Eq. (4). Also note that despite the fact that $SP_{item}$ and $SP_D$ are based on different "detection" events, they are compared in this paper; see also section 5.

For the example in (6), Figure 2 plots $P_D(n,r)$ for $n = 27$ (as for the sampling plan in Figure 1) and $n = 6$. $n = 6$ is the smallest sample size such that $P_D(n,r) \ge 0.2$ for all $r$ from the set $\{\lceil M/\bar{x} \rceil, \dots, N\}$ of diversion strategies.

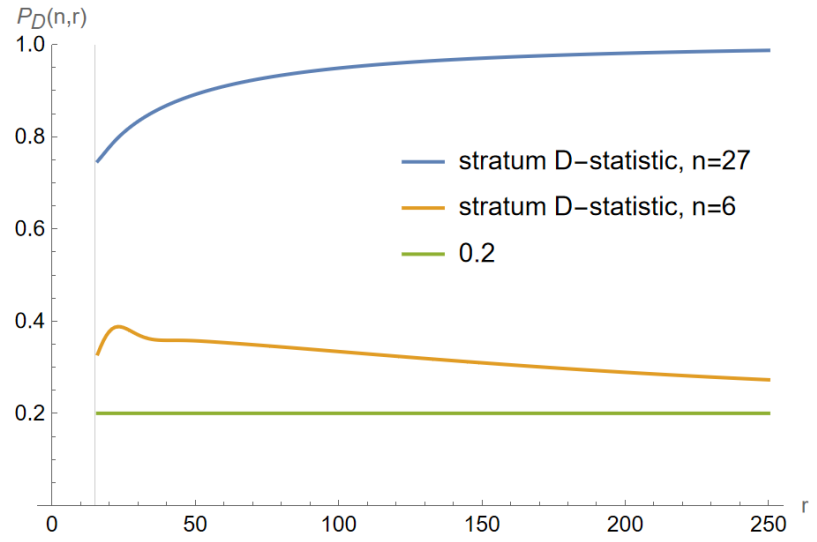


Figure 2: $P_D(27, r)$ as a function of the number $r$ of falsified items.

Figure 3 plots $DP(n, r)$ and $P_D(n, r)$ of $SP_{item}$ and $SP_D$ for $n = 27$.

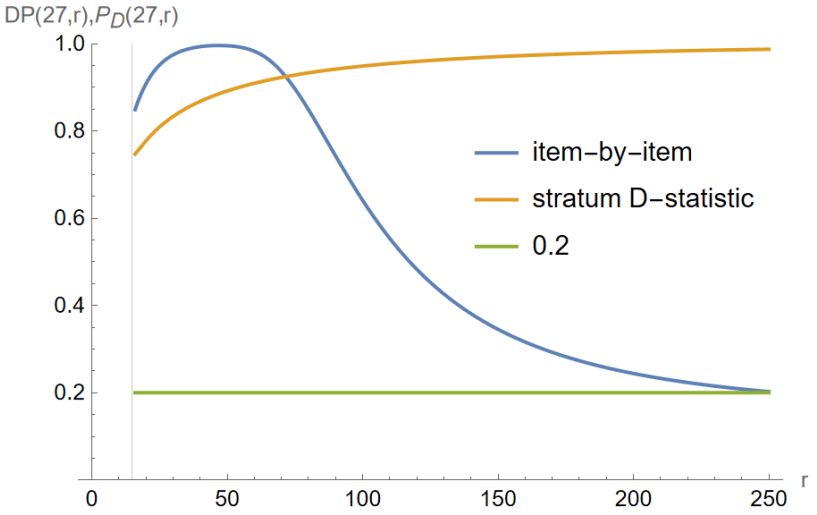


Figure 3: $DP(27, r)$ and $P_D(27, r)$ as a function of the number $r$ of falsified items.

Figure 3 shows that the sampling plan based on item-by-item tests is better at detecting large falsifications (gross defect, small values of $r$), and the sampling plan based on the relative stratum $D$-statistic is better at detecting small falsifications (bias defect, large values of $r$). The same is true if the sample size $n = 6$ is considered; see Figure 4.

# 4 Sampling plans based on the item-by-item tests and the relative stratum D-statistic ($SP_{item,D}$)

Now assume that the inspector performs during an inspection

1. the item-by-item tests (with clarification of any alarm), and then
2. – in case of not observing any significant item-by-item difference caused by a falsified item – in addition the $D$-statistic.

Thus, the inspector has more information available, and one could expect according to the motto "the more information the better" that this should bring something in terms of a higher probability of observing at least one significant difference for a falsified item or observing a significant value of the $D$- statistic.

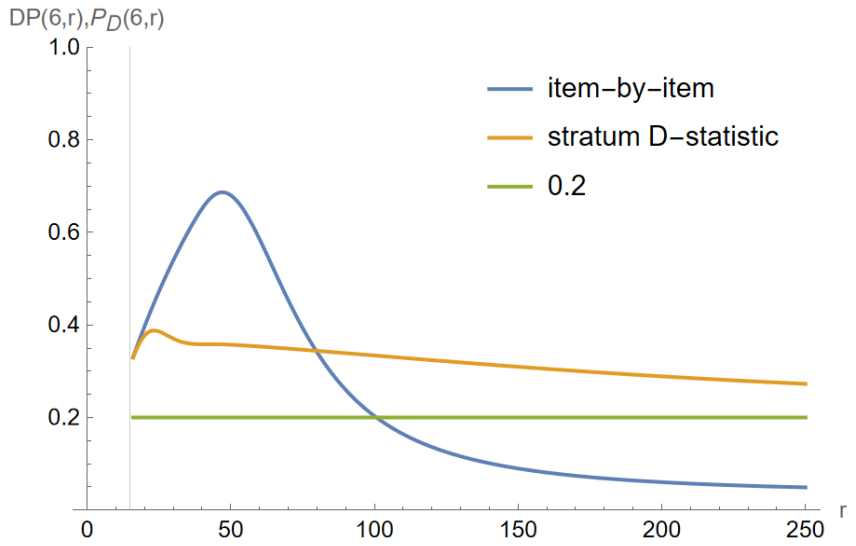


Figure 4: $DP(6,r)$ and $P_D(6,r)$ as a function of the number $r$ of falsified items.

Using the threshold $3\delta$ for the item-by-item tests and the threshold in Eq. (9) for the relative stratum difference statistic, the FAP is, by Eq. (10), given by

$$\begin{aligned} \alpha_3(n) &:= \mathbb{P}_{H_0}(\text{at least one false alarm}) \\ &= 1 - \mathbb{P}_{H_0}\left( D_1 \le 3\delta, \dots, D_n \le 3\delta, \frac{N}{n}\sum_{j=1}^{n} D_j \le k_D \right) \\ &\ge \alpha_1(n) = \alpha_2(n) \end{aligned} \tag{14}$$

because $\mathbb{P}(A \cap B) \le Min\big(\mathbb{P}(A), \mathbb{P}(B)\big)$ for any two events $A$ and $B$. Thus, the sampling plan $SP_{item,D}$ has a different baseline (with respect to the FAP) compared to $SP_{item}$ and $SP_D$. Thus, a comparison of all three sampling plans seems to be somewhat unreasonable at first sight; see Eq. (18) for a discussion.

Combining the events from Eqs. (4) and (12) we consider in $SP_{item,D}$ the event

$$\text{Combined detection} := \{detection\ event\} \cup \{\text{significant stratum } D - \text{statistic}\}, \tag{15}$$

in words: observing at least one significant (item-by-item) difference for a falsified item *or* observing a significant value of the relative stratum $D$-statistic when there is at least one falsified item among the sampled items.

The probability $P_{item,D}(n,r)$ of the event in Eq. (15) is, in analogy to Eqs. (5) and (13), given by

$$\begin{aligned} &P_{item,D}(n,r) = \mathbb{P}_{H_1}(\text{Combined detection}) \\ &:= \begin{cases} 1 - \sum_{i=r-(N-n)}^{Min(r,n)} \frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}}\, \mathbb{P}_{H_1}\left( Max(D_1,\dots,D_i) \le 3\delta, \frac{N}{n}\sum_{j=1}^{n} D_j \le k_D \right) \\ \qquad\qquad \text{for } r-(N-n) \ge 1 \\ 1 - \frac{\binom{N-r}{n}}{\binom{N}{n}} - \sum_{i=1}^{Min(r,n)} \frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}}\, \mathbb{P}_{H_1}\left( Max(D_1,\dots,D_i) \le 3\delta, \frac{N}{n}\sum_{j=1}^{n} D_j \le k_D \right) \\ \qquad\qquad \text{for } r-(N-n) \le 0\ . \end{cases} \end{aligned} \tag{16}$$

Again, we do not call $P_{item,D}(n,r)$ detection probability, because $P_{item,D}(n,r)$ is not based on the detection event in Eq. (15) while the current detection event is based on Eq. (4).

Because $\mathbb{P}(\bar{A} \cap \bar{B}) \le Min(\mathbb{P}(\bar{A}), \mathbb{P}(\bar{B}))$ for any two events $A$ and $B$, the inequality

$$\begin{aligned}\mathbb{P}(A \cup B) &= 1 - \mathbb{P}(\bar{A} \cap \bar{B}) \\ &\geq 1 - Min\big(\mathbb{P}(\bar{A}), \mathbb{P}(\bar{B})\big) = Max\big(1 - \mathbb{P}(\bar{A}), 1 - \mathbb{P}(\bar{B})\big) \\ &= Max\big(\mathbb{P}(A), \mathbb{P}(B)\big)\end{aligned}$$

holds, that further yields

$$P_{item,D}(n,r) \geq Max\big(DP(n,r), P_D(n,r)\big)\,, \tag{17}$$

if $A$ is the detection event of Eq. (4) and event $B$ is the "significant stratum $D$−statistic"-event of Eq. (12). Thus, $P_{item,D}(n,r)$ is always higher than the single probabilities $DP(n,r)$ and $P_D(n,r)$. Therefore, the statement "the more information the better" from the beginning of section 4 is indeed true in this context.

Performing $10^4$ simulations to estimate the probability $\mathbb{P}_{H_1}(\dots)$ in Eq. (16) leads for the example in (6) to the curves depicted in Figure 5. The time for producing the curves in Figure 5 is about 4 minutes and it is mainly due to the simulation of $\mathbb{P}_{H_1}(\dots)$ for the computation of $P_{item,D}(6,r)$.

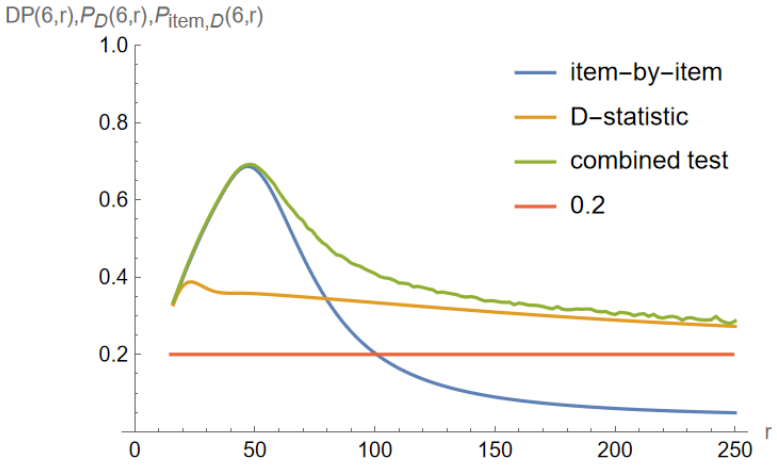


Figure 5: $DP(6,r)$, $P_D(6,r)$ and $P_{item,D}(6,r)$ as a function of the number $r$ of falsified items.

Figure 5 illustrates inequality (17): The combined test achieves the required DP of 0.2. Coming back to the starting point: In the current approach using $SP_{item}$, $n = 27$ items need to be verified to achieve the required 0.2 DP for all $r$ from the set $\{\lceil M/\bar{x} \rceil, \dots, N\}$ of diversion strategies; see Figure 1. If the combined test would be applied during inspections in the field, then the sample size could be reduced to (at least, see below) $n = 6$.

Practically, a conservative approach to find a smallest $n$ such that inequality (17) is fulfilled, is, to ensure that $Max\big(DP(n,r), P_D(n,r)\big)$ is at least the required DP for all $r \in \{\lceil M/\bar{x} \rceil, \dots, N\}$. Note that this $n$ can be found without using $P_{item,D}(n,r)$ that is based on time-consuming simulations.

However, if one desperately wants to determine $n$ using Eq. (16) instead of inequality (17), then an even greater reduction of the sample size $n$ is possible: For the example in (6) the curves $DP(n,r)$, $P_D(n,r)$ and $P_{item,D}(n,r)$ are depicted for $n = 5$ in Figure 6. While the combined test still achieves the required 0.2 DP, the D-statistic does not: $P_D(5{,}250) = 0.199$. While $n = 5$ is indeed a marginal difference to $n = 6$ and not worth the computational effort, other examples might result in more remarkable differences.

In Eq. (14) it is shown that a comparison of $SP_{item}$ and $SP_D$ with $SP_{item,D}$ is a bit unwarranted because the FAP of the $SP_{item,D}$ given by $\alpha_3(n) \geq \alpha_1(n) = \alpha_2(n)$ and so the tests performed have a different FAP. What is the FAP for the example in (6)? Eqs. (2) and (14) (with $10^5$ simulations for estimating $\alpha_3(n)$) lead to

$$\begin{aligned}\alpha_1(27) = \alpha_2(27) = 0.0358 \quad &\text{and} \quad \alpha_3(27) = 0.0678 \\ \alpha_1(6) = \alpha_2(6) = 0.0081 \quad &\text{and} \quad \alpha_3(6) = 0.0152\,.\end{aligned} \tag{18}$$

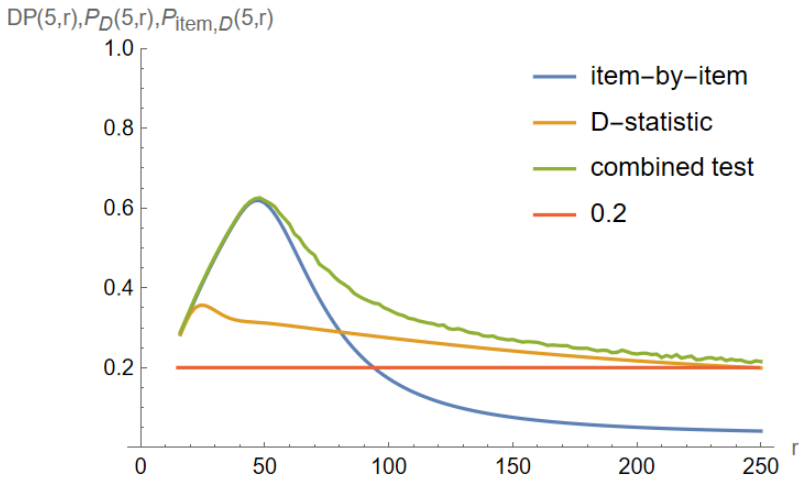


Figure 6: $DP(5,r)$, $P_D(5,r)$ and $P_{item,D}(5,r)$ as a function of the number $r$ of falsified items.

As expected, the FAPs are different $0.0358 < 0.0678$ for $n = 27$and $0.0081 < 0.0152$ for $n = 6$, but $\alpha_3(6) < \alpha_1(27) = \alpha_2(27)$. Thus, the fact that $\alpha_3(n) \geq \alpha_1(n) = \alpha_2(n)$ for $n = 6$ and 27 should not concern us too much, because the reduction of the sample size results in a smaller $\alpha_3(n)$ which is even smaller than the FAPs of $SP_{item}$: $\alpha_3(6) < \alpha_1(27) = \alpha_2(27)$.

Because $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ for any two events $A$ and $B$, the FAP $\alpha_3(n)$ fulfils by Eq. (2)

$$\alpha_3(n) = \mathbb{P}_{H_0}(\text{at least one false alarm}) \leq 2\,\alpha_1(n)\,,$$

and thus, cannot be larger than $2\,\alpha_1(n)$, i.e., $\alpha_1(n) \leq \alpha_3(n) \leq 2\,\alpha_1(n)$.

The sample size $n = 27$ used in Figure 1 is determined such that $DP(n,r) \geq 0.2$ for all $r$ from the set $\{\lceil M/\bar{x} \rceil, \ldots, N\}$ of diversion strategies. Currently the sample size used in a stratum is determined by the IAEA formula given by $n_{IAEA} = \left\lceil N\left(1 - \beta_{req}^{1/\lceil M/\bar{x} \rceil}\right)\right\rceil$, where $\lceil a \rceil$, $a \in \mathbb{R}$, is the smallest integer not less than $a$, and $\beta_{req}$ is the required non-detection probability. For the example in (6) and $\beta_{req} = 0.8$, we get $n_{IAEA} = 4$. Because the sample size $n = 5$ is the smallest sample size such that $P_{item,D}(n,r) \geq 0.2$ for all $r$ from the set $\{\lceil M/\bar{x} \rceil, \ldots, N\}$ (see Figure 6), for $n = 4$ even $P_{item,D}(n,r)$ drops below 0.2 for all $r$ at about 150 and larger, see Figure 7.

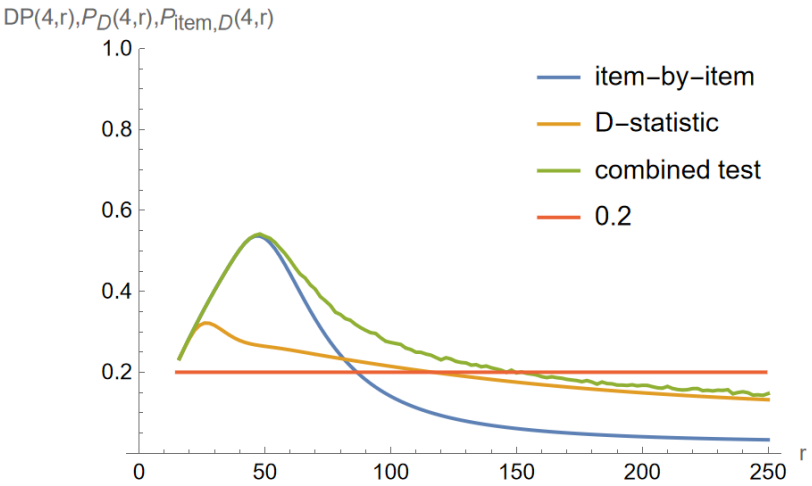


Figure 7: $DP(4,r)$, $P_D(4,r)$ and $P_{item,D}(4,r)$ as a function of the number $r$ of falsified items.

Figure 7 shows that if $n_{IAEA}$ is used as the sample size in $SP_{item,D}$, then the required DP is achieved at least for a wider range of diversion strategies compared to that of $SP_{item}$. This property is a nice benefit of using $SP_{item,D}$ instead of $SP_{item}$.

# 5 Summary and future work

If 1) the current detection event given by Eq. (4) is generalized to the "detection" event of Eq. (15) and if 2) a higher FAP with respect to $SP_{item}$ and $SP_D$ does not concerns too much, then the combined sampling plan $SP_{item,D}$ should be applied because it leads – at least for the numerical examples considered so far – to a remarkable reduction of the sample size (from $n = 27$ to $n = 5$) while the required detection probability is still achieved. In case that $n_{IAEA}$ is used as the sample size in $SP_{item,D}$, then the required DP is achieved for wider range of diversion strategies compared to that of $SP_{item}$.

In future work the results of this paper must be further investigated for a wider range of safeguards relevant input parameters $N, n, \delta, M$ and $\bar{x}$, and need to be generalized to the case that up to three measurement methods are applied in a stratum.

Some thoughts should be invested in the modified "detection" event of Eq. (15) and whether current IAEA approaches allow for such a modification.

Also, it must be elaborated of how to procced with a significant value of the relative stratum $D$-statistic and whether – as in the item-by-item test sampling plan – a further investigation will take place.

# 6 Acknowledgement

The work of Th. Krieger was supported by the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV), Germany, through the German Safeguards Support Programme to the IAEA under task D 1925/B.26.

# 7 References

[1] IAEA, Statistical Methods for Verification Sampling Plans (Safeguards Technical Report 381), Vienna: IAEA, 2017.

[2] T. Krieger, T. Burr and C. Norman, "Consequences of non-zero item variability on the IAEA's inspection sampling plans," in *Proceedings of the INMM 58th Annual Meeting*, Indian Wells, California USA, July 16-20, 2017.

[3] H.-P. Battenberg, Optimale Gegenstrategien bei Datenverifikations-Tests (Optimal counterstrategies for data verification tests), University of the German Federal Armed Forces Munich, Neubiberg, Germany, 1983.

[4] G. Piehlmeier, Spieltheoretische Untersuchungen von Problem der Datenverification (Game Theoretical Analyses of Data Verification Problems), Hamburg: Kovac, 1996.

[5] G. Casella und R. L. Berger, Statistical Inference, second Hrsg., Pacific Grove: Duxbury, 2002.

[6] V. K. Rohatgi, Statistical Inference, Mineola, N.Y.: Dover Publications Inc., 2003.

[7] IAEA, IAEA Safeguards Glossary, 2022 Edition (IAEA International Nuclear Verification Series No. 3, Rev. 1), Vienna: IAEA, 2022.