Screening Under Evidence Scarcity: A Coverage-Calibrated Theory of the Pre-Diligence Read

Daniel Ray Edgar
Finsider Labs

October 2026

Abstract

A screen run before a full Quality of Earnings engagement works with whatever evidence is on the table. I argue that its central danger is not the false alarm but the false all-clear: a clean result produced by checks that could not run. I formalise the screen as Bayesian evidence about a latent material defect, with coverage made explicit: a check contributes a likelihood ratio only if the evidence it needs is present, and an unassessable check contributes a ratio of exactly one. From this I derive the posterior after a clean screen in closed form, show that it equals the prior when nothing could be assessed, and compute the exact factor by which collapsing "insufficient data" into "clean" understates the odds of a defect. I give the log-likelihood threshold a clean screen must reach to support a target level of assurance, and I prove that the information a set of checks carries about the defect is monotone and submodular under conditional independence, so that greedy selection of the next document request is within a factor $1 - 1/e$ of the best fixed request list. I treat flags as evidence to be ranked by weight rather than counted, use a first-digit test as a worked example of a check whose power depends on sample size, derive the posterior threshold above which a screen should end a deal, and bound what a materiality floor can hide. The paper is theory and method. It reports no empirical results (its numerical examples are illustrative) and does not claim that a screen substitutes for diligence.

1 Introduction

A buyer considering an acquisition usually sees a company twice. The second look is the full Quality of Earnings engagement: weeks of work, complete ledgers, bank statements, payroll registers, management interviews. The first look happens earlier and with far less: a set of monthly statements, perhaps a general ledger export, and a few days in which to decide whether the deal deserves the cost of the second look. I call this first look the *screen*.

The seller knows more about the business than the buyer does, and the buyer knows it. This is the setting of Akerlof [1970]: when quality cannot be observed, the uninformed side must either discount every offer or invest in telling good from bad. A screen is the cheapest such investment, and the one most exposed to a particular failure. A screen is a list of checks, and each check needs evidence. When the evidence for a check is missing, the

check does not fail; it simply does not run. If the result is then summarised as "nothing found", the reader receives a statement that is true and nearly worthless, and cannot tell it apart from the same sentence produced by a screen in which every check ran and passed.

The claim of this paper is that this distinction can be made exact. I model the screen as a sequence of binary checks, each with a sensitivity and a false-alarm rate, each either assessable or not, and I track the posterior probability of a material defect. The model is deliberately small. It can say what a clean result is worth given what ran, which missing document would be most informative, and when the posterior is high enough that paying for diligence is a poor bet. It cannot supply the prior or the error rates, and it does not convert a screen into diligence. Sections 2 to 4 derive the worth of a clean screen; Sections 5 to 9 treat requests, flags, a digit test, the stopping decision and the materiality floor; the remaining sections give implications, a worked example, predictions and limitations.

2 The Model

Definition 2.1 (Deal, defect, screen)**.** A deal carries a latent variable $D \in \{0,1\}$, where $D = 1$ means that a material defect is present (reported earnings would not survive full diligence). The prior is $\pi = \mathbb{P}(D = 1) \in (0,1)$. A *screen* is a set of checks indexed by $[m] = \{1, \dots, m\}$. Check j is *assessable* if the evidence it needs is present. The random set $A \subseteq [m]$ of assessable checks is the *coverage*, and $U = [m] \setminus A$ is the set of unassessable checks. An assessable check returns $X_j \in \{0,1\}$, with $X_j = 1$ a *flag* and $X_j = 0$ a *pass*. Check j has sensitivity $s_j = \mathbb{P}(X_j = 1 \mid D = 1)$ and false-alarm rate $f_j = \mathbb{P}(X_j = 1 \mid D = 0)$. The screen is *clean* if every assessable check passes; I write clean for this event.

I impose three assumptions and return to each in Section 13.

(A1) Conditional independence. Given D and given $A = a$, the outcomes $(X_j)_{j \in a}$ are independent, with $\mathbb{P}(X_j = 1 \mid D = 1, A = a) = s_j$ and $\mathbb{P}(X_j = 1 \mid D = 0, A = a) = f_j$.
(A2) Ignorable coverage. The coverage is independent of the defect: $\mathbb{P}(A = a \mid D = 1) = \mathbb{P}(A = a \mid D = 0)$ for every $a \subseteq [m]$.
(A3) Informative checks. For every j, $0 < f_j < s_j < 1$.

Assumption (A1) is the one under which evidence multiplies, and the one most likely to fail in practice, because two checks that read the same ledger share its errors. Assumption (A2) says that evidence is missing for reasons unrelated to the defect; Remark 3.4 shows what changes when it fails. Assumption (A3) says that a flag is more likely with a defect than without, and keeps every logarithm finite.

For a set $a \subseteq [m]$ define the *miss probability* and the *quiet probability*

$$M_a = \prod_{j \in a} (1 - s_j), \qquad Q_a = \prod_{j \in a} (1 - f_j), \tag{1}$$

with the empty product equal to one. Under (A1), M_a is the probability that a defective deal produces a clean screen at coverage a, and Q_a is the probability that a sound deal does. Under (A3), $0 < M_a \le Q_a \le 1$, with $M_a < Q_a$ whenever a is nonempty.

3 What a Clean Screen Is Worth

Theorem 3.1 (Posterior after a clean screen)**.** *Assume* (A1) *to* (A3)*. For every coverage a with $\mathbb{P}(A = a) > 0$,*

$$p_a \;:=\; \mathbb{P}(D = 1 \mid A = a,\ \mathsf{clean}) \;=\; \frac{\pi M_a}{\pi M_a + (1-\pi)\, Q_a}. \tag{2}$$

Proof. By (A2) write $\mathbb{P}(A = a \mid D = d) = r_a > 0$ for both values of d. By (A1), $\mathbb{P}(A = a, \mathsf{clean} \mid D = 1) = r_a M_a$ and $\mathbb{P}(A = a, \mathsf{clean} \mid D = 0) = r_a Q_a$. Bayes' rule gives

$$\mathbb{P}(D = 1 \mid A = a, \mathsf{clean}) = \frac{\pi\, r_a M_a}{\pi\, r_a M_a + (1-\pi)\, r_a Q_a},$$

and r_a cancels. The denominator is positive because $Q_a > 0$ under (A3). □

Corollary 3.2 (What a clean result does and does not say)**.** *Under* (A1) *to* (A3)*:*

(a) If $a = \emptyset$ then $p_a = \pi$. A screen in which nothing could be assessed leaves the prior unchanged.

(b) If $a \subsetneq b$ then $p_b < p_a$. Each additional assessable check that passes strictly lowers the posterior; in particular $p_a < \pi$ for every nonempty a.

(c) For $a \subseteq b$, the posterior odds satisfy

$$\frac{\mathrm{odds}(p_a)}{\mathrm{odds}(p_b)} = \prod_{j \in b \setminus a} \frac{1 - f_j}{1 - s_j} \;\geq 1, \qquad \mathrm{odds}(p) = \frac{p}{1-p}. \tag{3}$$

(d) $p_a \geq \pi M_a / (1 - \pi(1 - M_a))$. No clean screen can lower the posterior below a floor fixed by the prior and the miss probability alone.

Proof. From (2), $\mathrm{odds}(p_a) = \frac{\pi}{1-\pi} \cdot \frac{M_a}{Q_a}$. (a) With $a = \emptyset$, $M_a = Q_a = 1$. (c) Dividing the odds at a by the odds at b leaves $\prod_{j \in b \setminus a} (1 - f_j)/(1 - s_j)$, and each factor exceeds one by (A3). (b) If $b \setminus a$ is nonempty the product in (c) is strictly greater than one, and $p \mapsto \mathrm{odds}(p)$ is strictly increasing. (d) Since $Q_a \leq 1$, the denominator of (2) is at most $\pi M_a + 1 - \pi$. □

Part (c) is the quantitative content of the thesis. Suppose a screen with coverage a is reported with the single word "clean", and the reader takes it, reasonably, as the clean result of the whole screen. The reader's odds of a defect are then too low by the factor $\prod_{j \in U} (1 - f_j)/(1 - s_j)$, which depends only on the checks that did not run. "Insufficient data" and "clean" are different outputs, and collapsing them overstates assurance by an amount that can be computed and so should be disclosed. The refusal to classify has long been treated as a legitimate third output with its own cost [Chow, 1970]; it is the correct output for an unassessable check.

Definition 3.3 (Coverage-weighted detection power)**.** The *detection power* of a screen at coverage a is $\mathrm{Pow}(a) = 1 - M_a = \mathbb{P}(\text{at least one flag} \mid D = 1,\ A = a)$, and its *false-alarm probability* is $\mathrm{Fa}(a) = 1 - Q_a$. With $\lambda_j = \log \frac{1-f_j}{1-s_j} > 0$ the weight of evidence of a pass on check j, the *evidence coverage* is $\kappa(a) = \sum_{j \in a} \lambda_j \big/ \sum_{j \in [m]} \lambda_j \in [0, 1]$.

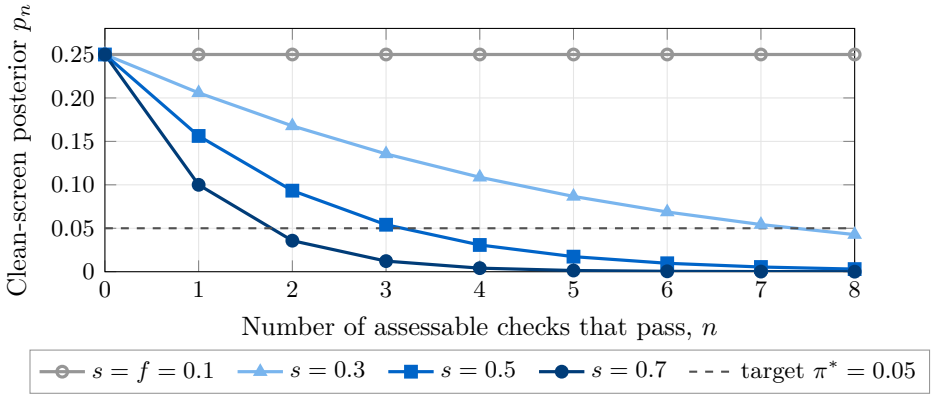


Figure 1: The clean-screen posterior of Theorem 3.1 for n assessable checks of equal sensitivity s and false-alarm rate $f = 0.1$, from a prior $\pi = 0.25$ (illustrative parameters). Every curve starts at the prior. The grey line is the boundary case $s = f$, excluded by (A3). By Corollary 4.3 the curve $s = 0.5$ first falls below the target at $n = 4$.

The identity in Definition 3.3 is immediate from (A1). I propose a reporting rule: *a clean result is never reported alone; it is reported as the triple* (clean, a, Pow(a)). By Corollary 3.2(d) the power alone bounds what the result can mean: with $w = \mathrm{Pow}(a)$, the posterior is at least $\pi(1-w)/(1-\pi w)$, whatever the false-alarm rates. Figure 1 plots (2) for checks of equal strength.

Remark 3.4 (When missing evidence is itself evidence)**.** If (A2) fails, write $\rho_a = \mathbb{P}(A = a \mid D = 1)/\mathbb{P}(A = a \mid D = 0)$. The proof of Theorem 3.1 then gives $\mathrm{odds}(p_a) = \frac{\pi}{1-\pi}\,\rho_a\, M_a/Q_a$: the pattern of missing evidence enters as one more likelihood ratio. If sellers with something to hide are slower to produce bank statements, then $\rho_a > 1$ for coverages that lack them and "insufficient data" is worse than uninformative. Assumption (A2) is thus the reading most favourable to the seller.

4 Likelihood-Ratio Form and the Assurance Threshold

The clean screen is a special case. For a general outcome vector, the natural statistic is the likelihood ratio, which Neyman and Pearson [1933] showed to be the basis of the most powerful test between two simple hypotheses.

Theorem 4.1 (Odds form)**.** *Assume* (A1) *to* (A3)*. For a coverage a and outcomes $x = (x_j)_{j\in a}$,*

$$\mathrm{odds}(\mathbb{P}(D = 1 \mid A = a,\ X_a = x)) = \frac{\pi}{1-\pi}\prod_{j\in a}\ell_j(x_j), \qquad \ell_j(1) = \frac{s_j}{f_j}, \quad \ell_j(0) = \frac{1-s_j}{1-f_j}. \tag{4}$$

Equivalently, the product may be taken over all of $[m]$ with every unassessable check contributing the factor 1*.*

Proof. As in Theorem 3.1, (A2) makes the factor $\mathbb{P}(A = a \mid D)$ cancel, and (A1) gives $\mathbb{P}(X_a = x \mid D = 1, A = a) = \prod_{j\in a} s_j^{x_j}(1-s_j)^{1-x_j}$ and the same expression with f_j under

$D = 0$. The ratio of the two is $\prod_{j \in a} \ell_j(x_j)$. A check outside a has no outcome and no factor, which is the same as a factor of one. □

Under (A3), $\ell_j(1) > 1 > \ell_j(0)$: a flag raises the odds, a pass lowers them, and an unassessable check does neither. It is silent, not reassuring.

Proposition 4.2 (Evidence needed for a target all-clear)**.** *Fix a target posterior $\pi^* \in (0, 1)$. Under* (A1) *to* (A3)*, a clean screen at coverage a satisfies $p_a \le \pi^*$ if and only if*

$$\sum_{j \in a} \lambda_j \;\ge\; \Lambda^*(\pi, \pi^*) := \log \frac{\pi\,(1-\pi^*)}{(1-\pi)\,\pi^*}, \qquad \lambda_j = \log \frac{1-f_j}{1-s_j}. \tag{5}$$

Proof. By Theorem 4.1, $\log \operatorname{odds}(p_a) = \log \frac{\pi}{1-\pi} - \sum_{j \in a} \lambda_j$. Since odds is strictly increasing, $p_a \le \pi^*$ if and only if $\log \frac{\pi}{1-\pi} - \sum_{j \in a} \lambda_j \le \log \frac{\pi^*}{1-\pi^*}$, which rearranges to (5). □

Corollary 4.3 (Minimum number of equal checks)**.** *If every check has the same pair (s, f) with $0 < f < s < 1$, and $\pi^* < \pi$, then a clean screen reaches $p \le \pi^*$ if and only if the number n of assessable checks satisfies*

$$n \;\ge\; n^* = \left\lceil \frac{\Lambda^*(\pi, \pi^*)}{\log\big((1-f)/(1-s)\big)} \right\rceil.$$

Proof. Here $\sum_{j \in a} \lambda_j = n\lambda$ with $\lambda = \log((1-f)/(1-s)) > 0$, and $\Lambda^* > 0$ because $\pi^* < \pi$. The least integer n with $n\lambda \ge \Lambda^*$ is $\lceil \Lambda^*/\lambda \rceil$. □

As an illustration, take $\pi = 0.25$, $\pi^* = 0.05$, $s = 0.5$, $f = 0.1$. Then $\Lambda^* = \log(19/3) \approx 1.846$ and $\lambda = \log 1.8 \approx 0.588$, so $n^* = \lceil 3.14 \rceil = 4$, in agreement with Figure 1. In the notation of Definition 3.3, the clean-screen odds are $\frac{\pi}{1-\pi} \exp\big(-\kappa(a)\Lambda_{[m]}\big)$ with $\Lambda_{[m]} = \sum_{j \in [m]} \lambda_j$: coverage scales the exponent, and a screen that reaches a target at full coverage falls short at coverage a by exactly $(1 - \kappa(a))\Lambda_{[m]}$ on the log-odds scale.

5 Which Document to Request Next

When a check is unassessable, the analyst can ask for the evidence it needs. Requests cost time and goodwill, so the order matters. I measure the value of a request by the expected reduction in uncertainty about D, using the entropy of Shannon [1948] in the way Lindley [1956] proposed for comparing experiments. In this section the coverage is chosen by the analyst, and a requested document is assumed to arrive whether or not the deal is defective, which is (A2) again.

For a finite random variable Y let $H(Y)$ be its entropy in bits, $H(Y \mid Z)$ the conditional entropy and $I(Y; Z) = H(Y) - H(Y \mid Z)$ the mutual information. Write $h(q) = -q \log_2 q - (1-q) \log_2 (1-q)$ for the binary entropy. For a set of checks $S \subseteq [m]$ write $X_S = (X_j)_{j \in S}$.

The *information gain* of a set of checks S is $F(S) = I(D; X_S) = H(D) - H(D \mid X_S)$, with $F(\emptyset) = 0$.

Lemma 5.1 (Gain of a single check)**.** *Under* (A1)*, with $q_j = \pi s_j + (1-\pi) f_j$ the marginal probability of a flag,*

$$F(\{j\}) = h(q_j) - \pi\, h(s_j) - (1-\pi)\, h(f_j).$$

Proof. By symmetry of mutual information, $I(D; X_j) = H(X_j) - H(X_j \mid D)$. The variable X_j is Bernoulli with parameter q_j, and given $D = 1$ or $D = 0$ it is Bernoulli with parameter s_j or f_j. □

Theorem 5.2 (Monotonicity and submodularity)**.** *Assume the outcomes $(X_j)_{j\in[m]}$ are conditionally independent given D. Then the information gain F is*

(i) monotone: $F(S) \le F(T)$ whenever $S \subseteq T \subseteq [m]$;

(ii) submodular: for $S \subseteq T \subseteq [m]$ and $j \notin T$, $F(S \cup \{j\}) - F(S) \;\ge\; F(T \cup \{j\}) - F(T)$.

Proof. For any S and $j \notin S$, the chain rule for mutual information gives

$$F(S \cup \{j\}) - F(S) = I(D; X_j \mid X_S) = H(X_j \mid X_S) - H(X_j \mid D, X_S).$$

(i) Conditional mutual information is nonnegative, so adding one check never lowers F; adding the elements of $T \setminus S$ one at a time gives $F(S) \le F(T)$. (ii) Conditional independence given D means that, given D, X_j is independent of X_S, so $H(X_j \mid D, X_S) = H(X_j \mid D)$, and likewise with T in place of S. Hence

$$\big[F(S \cup \{j\}) - F(S)\big] - \big[F(T \cup \{j\}) - F(T)\big] = H(X_j \mid X_S) - H(X_j \mid X_T).$$

Since $S \subseteq T$, the variables X_T include X_S, and conditioning on more variables cannot increase entropy, so the right side is nonnegative. □

The proof uses nothing about D being binary; it holds for any finite-valued latent state. That information gain is submodular when observations are conditionally independent given the latent variables is the result of Krause and Guestrin [2005], who used it for near-optimal selection of observations in graphical models; the argument above is a short direct proof. The optimisation guarantee rests on a classical theorem of Nemhauser et al. [1978]: if a set function on a finite ground set is nondecreasing and submodular and vanishes on the empty set, then the greedy algorithm, which repeatedly adds an element of largest marginal gain, returns after k steps a set whose value is at least $1 - (1 - 1/k)^k \ge 1 - 1/e$ times the maximum over all sets of at most k elements.

Corollary 5.3 (Greedy requests are near-optimal)**.** *Let the outcomes be conditionally independent given D, let $W \subseteq [m]$ be the set of checks that could be made assessable by a request, and let G_k be the set chosen by $k \le |W|$ greedy steps on F restricted to W. Then*

$$F(G_k) \;\ge\; \Big(1 - \frac{1}{e}\Big) \max_{S \subseteq W,\ |S| \le k} F(S).$$

Proof. By Theorem 5.2, F restricted to subsets of W is nondecreasing and submodular, and $F(\emptyset) = 0$. These are the hypotheses of the theorem of Nemhauser et al. [1978] quoted above. □

Three qualifications apply. First, the guarantee concerns a request list fixed in advance, with equal-cost requests, judged by expected information; it is not a statement about adaptive policies that choose each request after seeing the previous outcome, although the greedy rule

can be applied adaptively. Second, the result applies mid-screen: conditional on outcomes already observed, D has the posterior of Theorem 4.1 as its new prior and the remaining checks are still conditionally independent given D. Third, a document often unlocks several checks. If request r unlocks the set C_r, then $R \mapsto F\big(\bigcup_{r\in R} C_r\big)$ is again nondecreasing and submodular, because for $S \subseteq T$ and any B submodularity and monotonicity give $F(S\cup B)+F(T) \geq F(T\cup B)+F\big((S\cup B)\cap T\big) \geq F(T\cup B)+F(S)$; so the corollary applies to documents as well as to checks.

6 Flags: Ranking, Not Counting

The opposite error to the false all-clear is the screen that always finds something.

Proposition 6.1 (At least one false flag)**.** *Under* (A1)*, a sound deal screened at coverage a shows at least one flag with probability* $\mathrm{Fa}(a) = 1 - Q_a$*, and the expected number of flags is* $\sum_{j\in a} f_j$*. If the n assessable checks share a false-alarm rate α, the probability is* $1-(1-\alpha)^n$*.*

Proof. Given $D = 0$ the outcomes are independent with $\mathbb{P}(X_j = 0) = 1 - f_j$, so the probability of no flag is Q_a, and expectation is linear. □

With $\alpha = 0.05$ the probability of at least one flag on a sound company passes one half at $n = 14$ and reaches 0.64 at $n = 20$. A screen judged by whether it "found something" rewards adding checks regardless of their quality.

Proposition 6.2 (The posterior depends on the weight of flags, not their number)**.** *Under* (A1) *to* (A3)*, the posterior after outcomes x at coverage a is a strictly increasing function of the weight of evidence*

$$\omega(x) = \sum_{j\in a:\ x_j=1} \log\frac{s_j}{f_j} \;-\; \sum_{j\in a:\ x_j=0} \lambda_j,$$

and the number of flags is not a function of ω: an outcome with fewer flags can carry a higher posterior than an outcome with more.

Proof. By Theorem 4.1, log odds of the posterior equals $\log\frac{\pi}{1-\pi} + \omega(x)$, and odds is strictly increasing. For the second claim it suffices to give one instance. Take $\pi = 1/4$ and three assessable checks with (s, f) equal to $(0.6, 0.1)$, $(0.4, 0.2)$ and $(0.5, 0.25)$. If only the first flags, the likelihood ratio is $\frac{0.6}{0.1}\cdot\frac{0.6}{0.8}\cdot\frac{0.5}{0.75} = 3$ and the posterior is $1/2$. If the second and third flag and the first passes, the ratio is $\frac{0.4}{0.9}\cdot\frac{0.4}{0.2}\cdot\frac{0.5}{0.25} = \frac{16}{9}$ and the posterior is $16/43 \approx 0.372$. □

Many checks are themselves families of item-level tests: the same test applied to each account, month or counterparty, each with its own null hypothesis and p-value. Here the question is which items to surface, and the relevant error measure is the false discovery rate, the expected proportion of surfaced items whose null hypothesis is true. Benjamini and Hochberg [1995] proved that the following step-up procedure controls it. Order the p-values $p_{(1)} \leq \cdots \leq p_{(N)}$ of N hypotheses, let k be the largest i with $p_{(i)} \leq iq/N$, and reject the hypotheses with the k smallest p-values (none if no such i exists). If the test statistics are independent, the false discovery rate is at most $(N_0/N)\,q \leq q$, where N_0 is the number of true null hypotheses. Independence is part of the theorem as I use it, and item-level tests on a single ledger may not satisfy it. The procedure is itself a ranking rule: it surfaces the smallest p-values first and stops at a data-dependent depth.

7 A Digit Test as an Example Check

To see how a real check sits inside the model, consider the first-digit test. Benford [1938] observed that in many collections of numbers the leading significant digit $d \in \{1, \dots, 9\}$ occurs with frequency close to

$$b_d = \log_{10}\Big(1 + \frac{1}{d}\Big), \tag{6}$$

so that a leading 1 appears about 30.1% of the time and a leading 9 about 4.6%. Hill [1995] gave a statistical derivation of the law, and Nigrini [2012] developed its use in forensic accounting and auditing.

Proposition 7.1 (Benford's law is a probability distribution)**.** *For* $1 \le k \le 9$, $\sum_{d=1}^{k} b_d = \log_{10}(k+1)$. *In particular* $\sum_{d=1}^{9} b_d = 1$.

Proof. $b_d = \log_{10}(d+1) - \log_{10} d$, so the sum telescopes to $\log_{10}(k+1) - \log_{10} 1 = \log_{10}(k+1)$, and $\log_{10} 10 = 1$. Each b_d is positive. □

Given n amounts with first-digit counts $O_1, \dots, O_9$, the usual test statistic is Pearson's

$$\chi^2 = \sum_{d=1}^{9} \frac{(O_d - n b_d)^2}{n b_d},$$

which, if the first digits are independent draws from (6), is approximately chi-square with $9 - 1 = 8$ degrees of freedom for large n; the test flags at the 5% level when $\chi^2 > 15.51$. What the statistic does under any other digit distribution can be written exactly.

Proposition 7.2 (Mean of the digit statistic)**.** *Let the first digits be independent with* $\mathbb{P}(\textit{digit} = d) = b'_d$. *Then*

$$\mathbb{E}[\chi^2] = \sum_{d=1}^{9} \frac{b'_d(1 - b'_d)}{b_d} \;+\; n\,\delta^2, \qquad \delta^2 = \sum_{d=1}^{9} \frac{(b'_d - b_d)^2}{b_d}.$$

If $b' = b$ *the mean is* 8 *for every* n. *If* $b' \neq b$ *the mean grows linearly in* n.

Proof. O_d is binomial with parameters n and b'_d, so $\mathbb{E}[(O_d - n b_d)^2] = \mathrm{Var}(O_d) + (n b'_d - n b_d)^2 = n b'_d(1 - b'_d) + n^2 (b'_d - b_d)^2$. Dividing by $n b_d$ and summing gives the formula. If $b' = b$ then $\delta^2 = 0$ and the first sum is $\sum_d (1 - b_d) = 9 - 1 = 8$ by Proposition 7.1. □

The proposition makes the test's dependence on sample size transparent. A fixed departure $\delta^2 > 0$, however small and however innocent, will eventually be flagged as n grows, and a large departure in a short ledger will often not be flagged. If every digit were equally likely, for example, then $\delta^2 \approx 0.402$ and the mean $11.21 + 0.402\,n$ passes the critical value only at $n = 11$. The test's sensitivity and false-alarm rate are thus functions of n and of how non-Benford the company's legitimate data are. Many legitimate data sets do not follow the law: amounts confined to a narrow range, assigned numbers such as invoice or account identifiers, and prices set at fixed points are standard examples [Nigrini, 2012]. For a business that sells at a handful of fixed prices, the false-alarm rate of the digit test is close to one, the likelihood ratio s/f of a flag is close to one, and by Proposition 6.2 the flag should carry almost no weight. A departure from Benford's law is a prompt for inquiry, never proof of manipulation.

8 When Should a Screen End a Deal?

A posterior is useful only if it changes a decision. I use the simplest model in which the question has an answer.

Definition 8.1 (Two-stage decision)**.** After the screen the buyer holds a posterior p and either *stops*, with payoff 0, or *proceeds* to full diligence at cost $C > 0$. Full diligence detects a defect with probability $t \in [0, 1]$ when $D = 1$, independently of the screen's outcomes given D, and raises no false alarm when $D = 0$. If a defect is detected the buyer walks away, having spent C. Otherwise the deal closes: a sound deal is worth $G > C$, and a defect discovered only after close costs $L > 0$.

Theorem 8.2 (The stopping threshold)**.** *In the two-stage decision the expected payoff of proceeding at posterior p is*

$$V(p) = (1 - p)\, G - p\,(1 - t)\, L - C, \tag{7}$$

and proceeding is optimal if and only if

$$p \;\le\; p^{\dagger} = \frac{G - C}{G + (1 - t)\, L} \in (0, 1). \tag{8}$$

The threshold $p^{\dagger}$ is increasing in G and in t and decreasing in C and in L.

Proof. With probability $1-p$ the deal is sound and closes, for payoff $G-C$. With probability pt a defect is detected, for payoff $-C$. With probability $p(1-t)$ a defect is missed, for payoff $-C-L$. Summing gives (7). V is affine and strictly decreasing in p, with $V(p) \ge 0$ exactly when $p\,(G + (1-t)L) \le G - C$. The threshold is positive because $G > C$ and less than one because $C > 0$ and $(1-t)L \ge 0$. The numerator falls as C rises; the denominator rises with L and falls with t; and $\partial p^{\dagger}/\partial G = (C + (1-t)L)/(G + (1-t)L)^2 > 0$. □

A screen should end a deal, then, not when it finds a flag but when the weight of its flags carries the posterior above $p^{\dagger}$. The better the diligence that would follow (t near one), the more risk it is rational to carry into it.

Proposition 8.3 (Expected saving from screening)**.** *Let S be any stopping event determined by the screen's coverage and outcomes, with $\beta = \mathbb{P}(S \mid D = 1)$ and $\gamma = \mathbb{P}(S \mid D = 0)$, so that $\mathbb{P}(S) = \pi\beta + (1 - \pi)\gamma$. Relative to always proceeding, the policy "stop on S, otherwise proceed" changes the buyer's expected payoff by*

$$\Delta(S) = \underbrace{\mathbb{P}(S)\, C}_{\text{diligence not bought}} + \underbrace{\pi\, \beta\, (1 - t)\, L}_{\text{missed defects avoided}} - \underbrace{(1 - \pi)\, \gamma\, G}_{\text{good deals wrongly stopped}}, \tag{9}$$

before the cost of the screen itself. Among all stopping events, Δ is maximised by $S^* = \{p_X > p^{\dagger}\}$, where p_X is the posterior of Theorem 4.1, and $\Delta(S^*) \ge 0$.

Proof. Let $Y = (1 - D)G - D(1 - t)L - C$. By the argument of Theorem 8.2 and the conditional independence of diligence and screen given D, the expected payoff of proceeding

given the screen is $\mathbb{E}[Y \mid A, X_A] = V(p_X)$. Always proceeding earns $\mathbb{E}[Y]$; the policy earns $\mathbb{E}[Y\mathbf{1}_{S^c}]$. The difference is

$$\Delta(S) = -\mathbb{E}[Y\mathbf{1}_S] = \mathbb{P}(S)\,C + (1-t)L\,\mathbb{P}(S, D=1) - G\,\mathbb{P}(S, D=0),$$

which is (9). Also $\Delta(S) = -\mathbb{E}[V(p_X)\mathbf{1}_S]$ because S is determined by the screen. This is largest when S contains exactly the outcomes with $V(p_X) < 0$, that is $p_X > p^\dagger$, and it is then a sum of nonnegative terms. □

The saving is therefore not simply the diligence fees avoided: it is those fees, plus the losses on defective deals that diligence would have missed, minus the value of sound deals that the screen stopped. The last term is the cost of an over-eager screen.

The same model values a document request by its effect on the decision rather than on entropy. Let $W(p) = \max\{0, V(p)\}$ be the value of deciding optimally at posterior p. For an unassessable check j write $q = ps_j + (1-p)f_j$, $p^+ = ps_j/q$ and $p^- = p(1-s_j)/(1-q)$ for the probability of a flag and the posteriors after a flag and after a pass, so that $p^- < p < p^+$ and $qp^+ + (1-q)p^- = p$.

Proposition 8.4 (A request has decision value only if it can change the decision)**.** *The* decision value *of check j at posterior p, $\nu_j(p) = qW(p^+) + (1-q)W(p^-) - W(p)$, is nonnegative, and it is positive if and only if $p^- < p^\dagger < p^+$.*

Proof. V is affine, so $qV(p^+) + (1-q)V(p^-) = V(p)$, and $V(p^-) > V(p) > V(p^+)$. If $p \le p^\dagger$ then $W(p) = V(p)$ and $W(p^-) = V(p^-)$, so $\nu_j(p) = q(W(p^+) - V(p^+)) = q\max\{0, -V(p^+)\}$, which is positive exactly when $p^+ > p^\dagger$; here $p^- < p^\dagger$ holds automatically. If $p > p^\dagger$ then $W(p) = 0 = W(p^+)$ and $\nu_j(p) = (1-q)\max\{0, V(p^-)\}$, which is positive exactly when $p^- < p^\dagger$; here $p^+ > p^\dagger$ holds automatically. In both cases $0 < q < 1$ by (A3). □

Entropy and decision value can disagree: a check may be informative in the sense of Section 5 and still be worthless at the current posterior because neither of its outcomes crosses $p^\dagger$. This suggests a myopic three-way rule, drawn in Figure 2: if some available request has decision value above its cost, *request* the one with the largest value; otherwise *proceed* if $p \le p^\dagger$ and *stop* if $p > p^\dagger$. The rule has the two-threshold shape of the sequential tests of Wald [1945]. I do not claim it is the optimal sequential policy, which would require looking more than one request ahead.

9 The Materiality Floor as a Detection Threshold

Every screen has a size below which it does not surface an item. Auditing standards treat materiality as a matter of judgement about what would influence users of the statements [International Auditing and Assurance Standards Board, 2009]; a screen needs a mechanical rule. The rule I analyse is a floor proportional to revenue with an absolute minimum:

$$\tau(R) = \max\{0.01\,R,\ 1000\}, \tag{10}$$

where $R > 0$ is revenue in dollars. A candidate item with absolute amount below $\tau(R)$ is not surfaced.

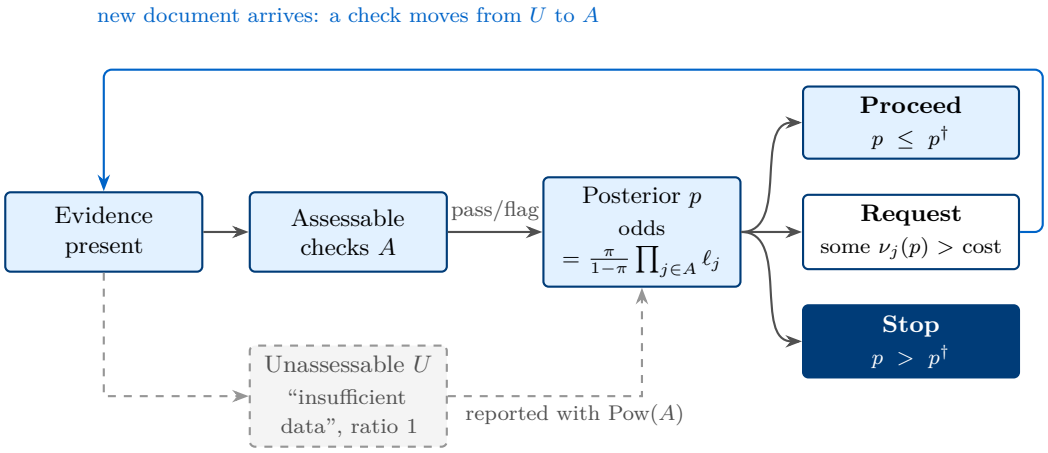


Figure 2: The screen as a pipeline. Only assessable checks contribute likelihood ratios; unassessable checks contribute a ratio of one and are reported as "insufficient data" with the detection power of what did run. The posterior is compared with $p^\dagger$ (Theorem 8.2); a request is made when some unassessable check has decision value (Proposition 8.4) above its cost.

Proposition 9.1 (What a floor can hide)**.** *Let the candidate items have amounts $c_1, \ldots, c_K$, let $\mathcal{E} = \{i : |c_i| < \tau(R)\}$ be the excluded set, $N = |\mathcal{E}|$ and $T = \sum_{i \in \mathcal{E}} |c_i|$. Then:*

(i) $\left|\sum_{i \in \mathcal{E}} c_i\right| \leq T \leq N\,\tau(R)$, with $T < N\tau(R)$ when $N \geq 1$;

(ii) the bound is sharp: for every $N \geq 1$ and $\varepsilon > 0$ there are amounts with $T > N\tau(R) - \varepsilon$.

Proof. (i) The first inequality is the triangle inequality; the second sums N terms each strictly less than $\tau(R)$. (ii) Take $|c_i| = \tau(R) - \varepsilon/(2N)$ for N items (with ε small enough that this is positive), giving $T = N\tau(R) - \varepsilon/2$. □

The floor bounds each omitted item but not their sum: for a company with revenue above $100,000, a hundred excluded items could in principle total nearly the whole of revenue. The floor alone cannot support the sentence "nothing material was omitted"; the pair (N, T) can, because it replaces the worst-case bound $N\tau$ by the actual total. Reporting (N, T) is the coverage disclosure for the dimension of size.

10 Implications for Practice

Remark 10.1 (Deterministic core, probabilistic periphery: a design invariant)**.** Let $\mathcal{R}$ be the rows parsed deterministically from ledger data, let $\mathcal{Z}$ be figures proposed by an AI extraction layer, and let $\mathcal{Z}_{\mathrm{v}} \subseteq \mathcal{Z}$ be those a human reviewer has verified. Suppose every headline metric has the form $g(\mathcal{R} \cup \mathcal{Z}_{\mathrm{v}})$. Then two extraction outputs $\mathcal{Z}$ and $\mathcal{Z}'$ with the same verified part give identical headline metrics, whatever errors their unverified parts contain. This is true by non-dependence and is not a deep statement; its value is as an invariant that a system either maintains or does not. The extraction layer can still raise coverage: a proposed figure, once verified, may make a check assessable.

j	Check	s_j	f_j	$\frac{1-s_j}{1-f_j}$	λ_j	$\log \frac{s_j}{f_j}$	Partial
1	Revenue to bank deposits	0.60	0.10	4/9	0.811	1.792	no
2	Accruals against cash flow	0.50	0.20	5/8	0.470	0.916	yes
3	Margin discontinuity	0.40	0.20	3/4	0.288	0.693	yes
4	First-digit test	0.20	0.05	16/19	0.172	1.386	yes
5	Payroll to headcount	0.50	0.10	5/9	0.588	1.609	no
6	Add-back support	0.50	0.25	2/3	0.405	0.693	no
		M_a	Q_a	Pow(a)	Fa(a)	$\kappa(a)$	p_a
No coverage, $a = \emptyset$		1	1	0	0	0	$1/4 = 0.2500$
Partial coverage, $a = \{2,3,4\}$		0.24	0.608	0.760	0.392	0.340	$5/43 \approx 0.1163$
Full coverage, $a = [6]$		0.024	0.36936	0.976	0.631	1	$100/4717 \approx 0.0212$

Table 1: An illustrative six-check screen with prior $\pi = 1/4$. Upper block: each check's sensitivity, false-alarm rate, pass likelihood ratio, pass weight λ_j, flag weight (natural logarithms) and assessability in the partial-coverage case. Lower block: the clean-screen quantities of Theorem 3.1 and Definition 3.3. All parameters are invented.

The invariant matters because readers over-trust automated output, a pattern Parasuraman and Riley [1997] describe as misuse of automation, and because the stated confidence of a learned model need not match its accuracy [Guo et al., 2017].

I refer to Finsider's product only where the theory touches a design choice. Finsider computes a deterministic core from ledger data; any figure extracted by an AI layer is kept separate, labelled as needing verification, and does not feed headline metrics until verified (Remark 10.1). Its screening materiality floor is one per cent of mapped revenue and never below USD 1,000, which is (10), and the count and total of candidates excluded under it are reported, as Proposition 9.1 requires. It keeps immutable original source files and traces reported figures to source rows, the kind of provenance studied by Buneman et al. [2001], so that a flag can be inspected, not merely believed. A company's results are held behind a release gate until verification passes, with human reviewers in the loop.

11 A Worked Example

Table 1 sets out an illustrative screen with six checks. The sensitivities and false-alarm rates are invented for arithmetic convenience and are not estimates of anything; the checks are named after familiar ideas, among them the divergence of accruals from cash flows studied by Sloan [1996] and Dechow and Dichev [2002] and ratio screens of the kind proposed by Beneish [1999]. The prior is $\pi = 1/4$. In the partial-coverage case the analyst holds only monthly statements and a ledger export, so checks 2, 3 and 4 are assessable; checks 1, 5 and 6 need bank statements, a payroll register and invoices.

Clean posteriors. At partial coverage $M_a = 0.5 \cdot 0.6 \cdot 0.8 = 0.24$ and $Q_a = 0.8 \cdot 0.8 \cdot 0.95 = 0.608$, so $p_a = \frac{0.06}{0.06+0.456} = \frac{5}{43}$. At full coverage $M = 0.024$ and $Q = 0.36936$, so $p = \frac{0.006}{0.006+0.27702} = \frac{100}{4717}$. The same word, "clean", describes a posterior of 11.6% in one case and 2.1% in the other. The odds are 5/38 and 100/4617, and their ratio $\frac{4617}{760} = 6.075$ is

exactly $\frac{9}{4} \cdot \frac{9}{5} \cdot \frac{3}{2} = \frac{243}{40}$, the product in (3) over the three checks that did not run. The floor of Corollary 3.2(d) at partial coverage is $0.06/0.81 \approx 0.074$.

False alarms. At full coverage a sound company shows at least one flag with probability $1 - 0.36936 \approx 0.631$. A single flag on check 1 with all else clean gives a posterior of about 0.226; a single flag on check 3 gives about 0.055.

The next request. Starting from the clean partial screen, the prior for the remaining checks is $5/43$, with entropy $h(5/43) \approx 0.5186$ bits. By Lemma 5.1 the gains of checks 1, 5 and 6 are approximately 0.1025, 0.0703 and 0.0210 bits, so the greedy first request is the bank statements. The gains of the pairs $\{1, 5\}$, $\{1, 6\}$ and $\{5, 6\}$ are approximately 0.1570, 0.1192 and 0.0883 bits, so the greedy pair $\{1, 5\}$ is here also the best pair, and check 5 adds 0.0703 bits alone but only 0.0545 after check 1, as Theorem 5.2 requires.

The decision. Take, in millions of dollars, $G = 2$, $L = 20$, $C = 0.3$ and $t = 0.8$. Then $p^\dagger = \frac{1.7}{2+4} = \frac{17}{60} \approx 0.283$ and $V(1/4) = 1.5 - 1 - 0.3 = 0.2$, so without a screen the buyer proceeds. At the clean partial posterior $5/43$, a flag on check 1 would give $15/34 \approx 0.441$ and a flag on check 5 would give $25/63 \approx 0.397$, both above $p^\dagger$, so by Proposition 8.4 each request has positive decision value. A flag on check 6 would give $5/24 \approx 0.208 < p^\dagger$: the invoices carry information and no decision value at this posterior. With the optimal stopping rule at full coverage, direct enumeration of the 64 outcomes gives $\beta = 0.78$ and $\gamma = 0.125335$, hence $\mathbb{P}(S^*) = 0.28900125$ and, by (9), $\Delta = 0.0867 + 0.78 - 0.1880 = 0.6787$: the screen raises the expected value of the deal process from 0.2 to about 0.879, before its own cost. At the partial coverage $\{2, 3, 4\}$, $\Delta = 0.339$. These figures illustrate the formulas and nothing else.

12 Falsifiable Predictions

A study with access to screens and their later diligence outcomes could test the following four claims.

Prediction 1 (Coverage predicts false all-clears)**.** Among screens that return a clean result, the rate at which full diligence later finds a material defect decreases as the screen's detection power $\mathrm{Pow}(a)$ increases. *Refuted if* clean screens with recorded coverage show no negative association between detection power and the later defect rate.

Prediction 2 (Dependence shows up as overconfidence)**.** Posteriors computed from (4) with error rates estimated on past deals will be overconfident at the extremes when checks share source data, and close to calibrated when checks draw on distinct evidence. *Refuted if* calibration error is no larger for screens whose checks share a source than for screens whose checks do not.

Prediction 3 (Weight beats count)**.** Ranking deals by the weight of evidence $\omega(x)$ separates deals that fail diligence from deals that pass better than ranking by the number of flags. *Refuted if* the flag count discriminates as well as or better than the weight, measured by area under the receiver operating characteristic curve on held-out deals.

Prediction 4 (The digit test over-flags structured ledgers)**.** On ledgers of sound companies whose amounts are dominated by fixed prices, assigned numbers or narrow ranges, the

first-digit test flags far more often than its nominal level, and the excess grows with ledger size as Proposition 7.2 implies. *Refuted if* the flag rate on such ledgers stays at the nominal level as the number of entries grows.

13 Limitations

Conditional independence. Assumption (A1) carries Theorems 3.1, 4.1 and 5.2. Checks that read the same ledger are typically positively dependent given D; multiplying their likelihood ratios then double-counts evidence, and I expect the clean posterior of (2) to be too low. Grouping dependent checks into one composite check with its own (s, f) restores the assumption at the cost of resolution.

The prior. Every posterior here is proportional, in odds, to the prior odds, and the model does not supply π. A base rate from comparable deals is defensible; a prior chosen to make the output reassuring is not, and the model cannot tell them apart. Leaving the prior unstated is no remedy: neglect of prior probabilities is among the best documented errors of intuitive judgement [Tversky and Kahneman, 1974].

Sensitivities are not known constants. The pairs (s_j, f_j) vary with company size, sector, ledger quality and, for statistical checks, sample size. They must be estimated, with uncertainty, from deals whose diligence outcomes are known, and such samples are small and selected: deals stopped at the screen never reveal their true state. A binary defect is also a simplification, and (A2) is a benchmark whose correction ρ_a in Remark 3.4 is harder to estimate than any s_j.

The decision model is minimal. Theorem 8.2 assumes one binary defect, a single detection probability, no false alarms in diligence, and risk-neutral payoffs with no renegotiation, whereas real diligence reprices deals more often than it ends them. Corollary 5.3 covers equal-cost, non-adaptive requests and an entropy objective only.

A screen is not diligence. Nothing in this paper shows that a clean screen, at any coverage, establishes the quality of earnings. The paper shows how much a clean screen moves a stated prior under stated assumptions, and that the answer can be zero.

14 Conclusion

A screen's most dangerous output is its most comfortable one. The worth of a clean result is a closed-form function of the prior and of the checks that actually ran; it equals the prior when none ran; and reporting "clean" without coverage understates the odds of a defect by a factor that depends only on the checks that did not run. The practical upshot is modest: a pre-diligence read should say what it could not check as precisely as it says what it found.

References

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

Messod D. Beneish. The detection of earnings manipulation. *Financial Analysts Journal*, 55 (5):24–36, 1999.

Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *Database Theory: ICDT 2001*, volume 1973 of *Lecture Notes in Computer Science*, pages 316–330. Springer, 2001.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Patricia M. Dechow and Ilia D. Dichev. The quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review*, 77(s-1):35–59, 2002.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

Theodore P. Hill. A statistical derivation of the significant-digit law. *Statistical Science*, 10 (4):354–363, 1995.

International Auditing and Assurance Standards Board. International standard on auditing 320: Materiality in planning and performing an audit. IFAC, 2009.

Andreas Krause and Carlos Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 324–331, 2005.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions, I. *Mathematical Programming*, 14(1): 265–294, 1978.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231: 289–337, 1933.

Mark J. Nigrini. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. John Wiley & Sons, Hoboken, NJ, 2012.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Richard G. Sloan. Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, 71(3):289–315, 1996.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.