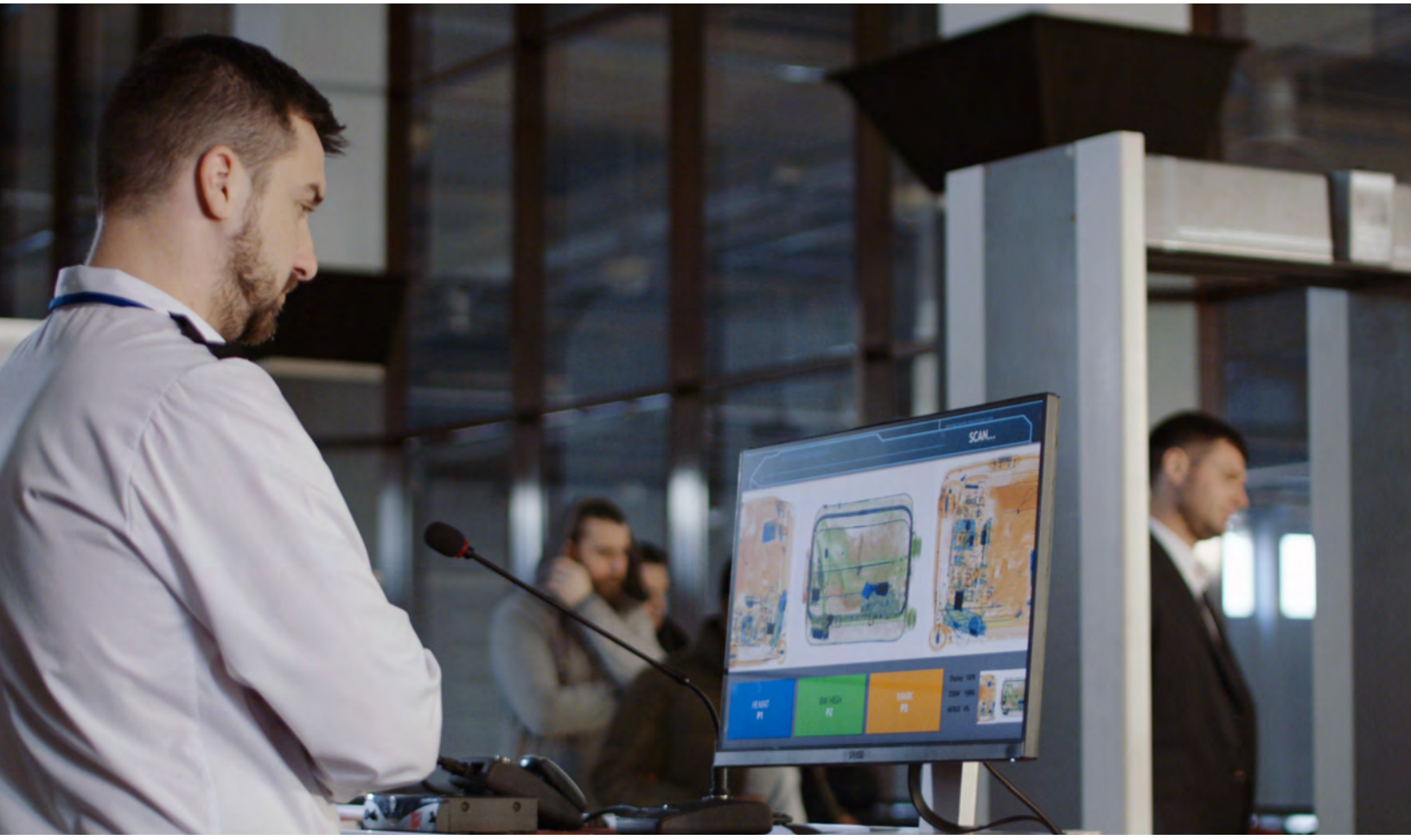
The Threat Detection System That Cried Wolf: Reconciling Developers with Operators[1]

Shelley M. Cazares

Both the Department of Defense (DoD) and the Department of Homeland Security (DHS) use threat detection systems, such as airplane cargo screeners and counter–improvised-explosive-device (IED) systems. These systems may perform well during testing but "cry wolf" in the field (i.e., generate false alarms when true threats are not present). As a result, operators can lose faith in the systems—ignoring them or even turning them off and taking the chance that a true threat will not occur. This paper reviews statistical concepts to reconcile the performance metrics that summarize a developer's view of a system during testing with the metrics that describe an operator's view of the system during real-world missions. Program managers can still make use of systems that cry wolf by arranging them into a tiered system that performs better than each individual system alone.



[1] The original article of the same title was published in *Defense Acquisition Research Journal*, January 2017, https://doi.org/10.22594/dau.16-749.24.01. The original article illustrates how a PM can make use of a system that frequently cries wolf by incorporating it into a tiered system that, overall, exhibits better performance than each individual system does alone.

Introduction

DoD and DHS operate counter-mine systems, counter-IED systems, airplane cargo screening systems, and other threat detection systems, all of which share a common purpose: to detect potential threats among clutter.

Threat detection systems are often assessed based on their Probability of Detection (P_d) and Probability of False Alarm (P_{fa}) (Urkowitz 1967). P_d describes the fraction of true threats for which the system correctly declares an alarm. Conversely, P_{fa} describes the fraction of true clutter (true nonthreats) for which the system *in*correctly declares an alarm—a false alarm. A perfect system will exhibit a P_d of 1 and a P_{fa} of 0. P_d and P_{fa} are defined in Table 1.

While the Probability of Detection and the Probability of False Alarm summarize how much of the truth causes an alarm, Positive Predictive Value and Negative Predictive Value summarize how many alarms turn out to be true.

Table 1. Definitions of Common Metrics Used to Assess the Performance of Threat Detection Systems

Metric	Definition	Perspective
Probability of Detection (P_d)	The fraction of all items containing a true threat for which the system correctly declared an alarm	Developer
Probability of False Alarm (P_{fa})	The fraction of all items not containing a true threat for which the system incorrectly declared an alarm	Developer
Positive Predictive Value (*PPV*)	The fraction of all items causing an alarm that did end up containing a true threat	Operator
Negative Predictive Value (*NPV*)	The fraction of all items not causing an alarm that did not end up containing a true threat	Operator
Prevalence (*Prev*)	The fraction of items that contained a true threat (regardless of whether the system declared an alarm)	Not applicable

Threat detection systems with good P_d and P_{fa} performance metrics are not always well received by system operators, because some systems may "cry wolf," generating false alarms when true threats are not present. As a result, operators may lose faith in the systems, delaying their response to alarms (Getty et al. 1995) or ignoring them altogether (Bliss et al. 1995), potentially leading to disastrous consequences. This issue has arisen in military, national security, and civilian scenarios (Cushman 1987; Stuart 1987; Oldham 2006).

This issue often stems from an inappropriate choice of metrics—P_d and P_{fa}—used to assess the system's performance during testing. While P_d and P_{fa}

encapsulate the *developer's* perspective of the system's performance, these metrics do not encapsulate the *operator's* perspective. The operator's view can be better summarized with other metrics, namely Positive Predictive Value (*PPV*) and Negative Predictive Value (*NPV*) (Altman and Bland 1994). *PPV* describes the fraction of all alarms that correctly turn out to be true threats—a measure of how often the system does not cry wolf. Similarly, *NPV* describes the fraction of all *lack* of alarms that correctly turn out to be true clutter. From the operator's perspective, a perfect system will have *PPV* and *NPV* values equal to 1. *PPV* and *NPV* are also defined in Table 1.

Interestingly enough, the same threat detection system that satisfies the developer's desire to detect as much truth as possible can also disappoint the operator by crying wolf too often (Scheaffer and McClave 1995). A system can exhibit excellent P_d and P_{fa} values, while also exhibiting a poor *PPV* value. Unfortunately, low *PPV* values naturally occur when the Prevalence (*Prev*) of true threat among true clutter is extremely low (Parasuraman 1997; Scheaffer and McClave 1995), as is often the case in defense and homeland security scenarios. As summarized in Table 1, *Prev* is a measure of how widespread or common the true threat is. A *Prev* of 1 indicates a true threat is always present, while a *Prev* of 0 indicates a true threat is never present. As we shall see, a low *Prev* can lead to a discrepancy in how developers and operators view the performance of threat detection systems in DoD and DHS.

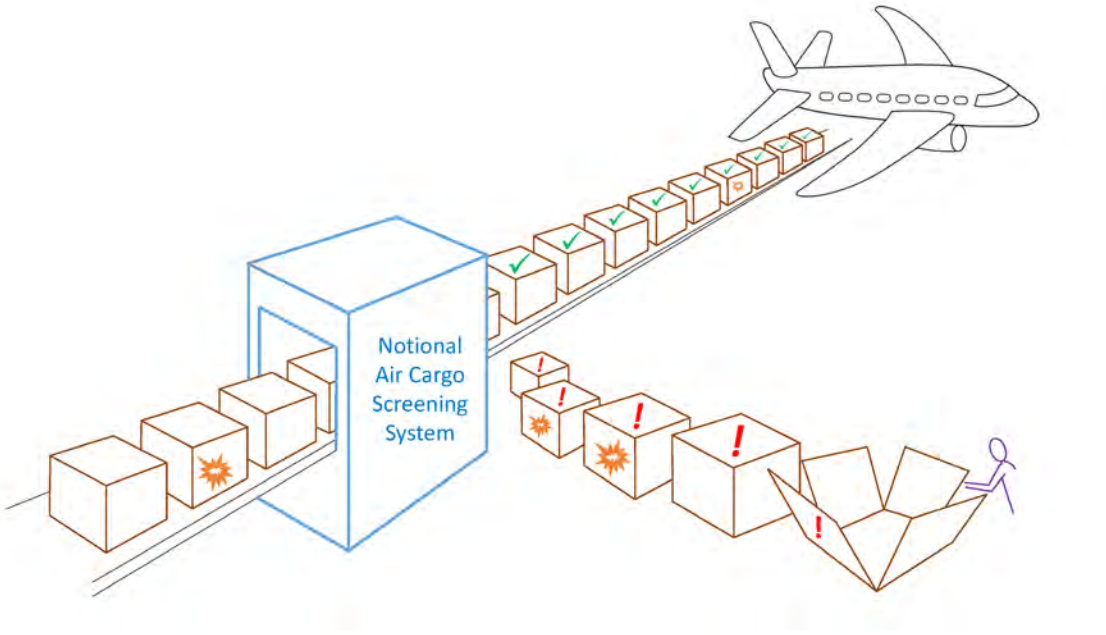
In the following sections, I reconcile the performance metrics used to quantify the developer's versus operator's views of threat detection systems. Although these concepts are already well known within the statistics and human factors communities, they are not often immediately understood in DoD and DHS science and technology acquisition communities. This review is intended for program managers (PMs) of threat detection systems in DoD and DHS.

Testing a Threat Detection System

Consider the notional air cargo screening system in Figure 1. The purpose of this notional system is to detect explosive threats packed inside items that are about to be loaded into the cargo hold of an airplane. To determine how well this system meets capability requirements, its performance must be quantified. A large number of items are input into the system, and each item's ground truth (whether the item contained a true threat) is compared to the system's output (whether the system declared an alarm). The items represent those that the system would likely encounter in an operational setting. At the end of the test, the following items are counted:

- True Positive (*TP*), an item containing a true threat for which the system correctly declared an alarm;
- False Positive (*FP*), an item *not* containing a true threat for which the system *in*correctly declared an alarm (a Type I error);

- False Negative (*FN*), an item containing true threat for which the system *in*correctly did *not* declare an alarm (a Type II error); and
- True Negative (*TN*), an item *not* containing a true threat for which the system correctly did *not* declare an alarm.



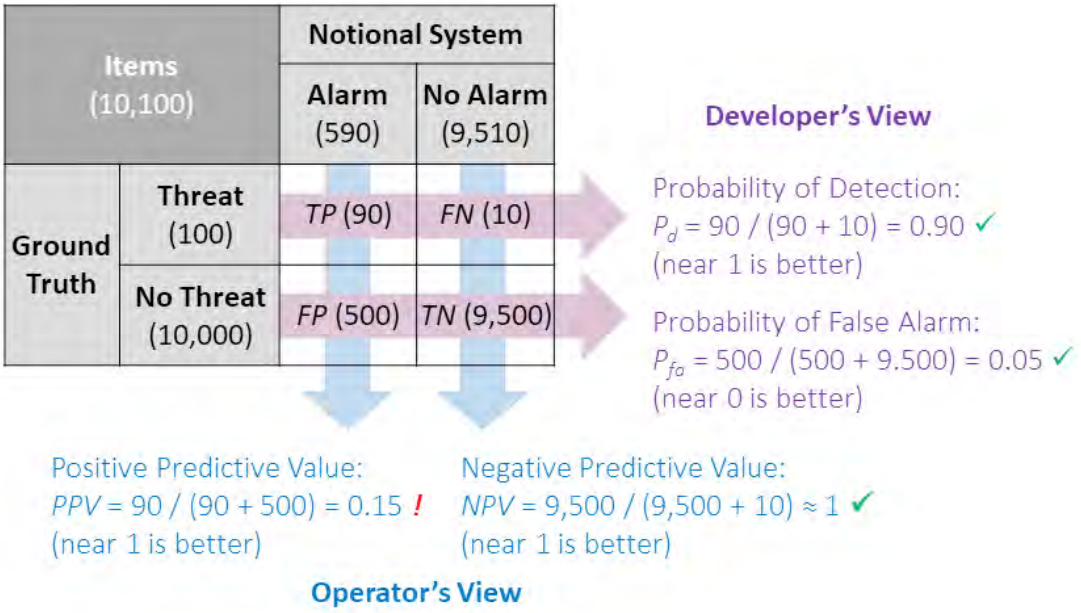
Note: A set of predefined, discrete items (small brown boxes) are presented to the system one at a time. Some items contain a true threat (orange star) among clutter, while other items contain clutter only (no orange star). For each item, the system declares either one or zero alarms. All items for which the system declares an alarm (red exclamation point) are further examined manually by trained personnel (purple figure). In contrast, all items for which the system does not declare an alarm (green checkmark) are left unexamined and loaded directly onto the airplane.

Figure 1. Notional Air Cargo Screening System

As shown in Figure 2, a total of 10,100 items passed through the notional air cargo screening system. One hundred items contained a true threat, while 10,000 items did not. The system declared an alarm for 590 items and did not declare an alarm for 9,510 items. Comparing the items' ground truth to the system's alarms (or lack thereof), there were 90 *TP*s, 10 *FN*s, 500 *FP*s, and 9,500 *TN*s.

Developer's View: P_d and P_{fa}

A PM must consider how much of the truth the threat detection system is able to identify. This can be done by considering two questions: Of those items that contain a true threat, for what fraction does the system correctly declare an alarm? And of those items that do *not* contain a true threat, for what fraction does the system *in*correctly declare an alarm? These questions often guide developers during the research and development phase of a threat detection system.



Note: This 2 × 2 matrix tabulates the number of *TP*, *FN*, *FP*, and *TN* items processed by the system. P_d and P_{fa} summarize the developers' view of the system's performance, while *PPV* and *NPV* summarize the operators' view. In this notional example, the low *PPV* of 0.15 indicates a poor operator experience (the system often cries wolf, since only 15 percent of alarms turn out to be true threats) even though the good P_d and P_{fa} are well received by developers.

Figure 2. 2 × 2 Confusion Matrix of a Notional Air Cargo Screening System

P_d and P_{fa} can be easily calculated from the confusion matrix to answer these questions. From a developer's perspective, the notional air cargo screening system exhibits good performance:[2]

$$P_d = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0.90 \text{ (compared to 1 for a perfect system)} \quad (1)$$

$$P_{fa} = \frac{FP}{FP + TN} = \frac{500}{500 + 9{,}500} = 0.05 \text{ (compared to 0 for a perfect system).} \quad (2)$$

Equation 1 shows that, of all items that contained a true threat (*TP* + *FN* = 90 + 10 = 100), a large subset (*TP* = 90) correctly caused an alarm. These counts resulted in P_d = 0.90, close to the value of 1 that would be exhibited by a perfect system.[3] Based on this P_d value, the PM can conclude that 90 percent of items

[2] PMs must determine what constitutes a "good" performance. For some systems operating in some scenarios, P_d = 0.90 is considered good, since only 10 FNs out of 100 true threats is considered an acceptable risk. In other cases, P_d = 0.90 is not acceptable. Appropriately setting a system's capability requirements calls for a frank assessment of the likelihood and consequences of *FN*s versus *FP*s and is beyond the scope of this paper.

that contained a true threat correctly caused an alarm, which may (or may not) be considered acceptable within the capability requirements for the system. Furthermore, Equation 2 shows that, of all items that did *not* contain a true threat ($FP + TN = 500 + 9{,}500 = 10{,}000$), only a small subset ($FP = 500$) caused a false alarm. These counts led to $P_{fa} = 0.05$, close to the 0 value that would be exhibited by a perfect system. In other words, only 5 percent of items that did *not* contain a true threat caused a false alarm.

Operator's View: *PPV* and *NPV*

The PM must also anticipate the operator's view of the threat detection system. One way to do this is to answer the following questions: Of those items that caused an alarm, what fraction turned out to contain a true threat (i.e., what fraction of alarms turned out *not* to be false)? And of those items that did *not* cause an alarm, what fraction turned out *not* to contain a true threat? On the surface, these questions seem similar to those posed previously for P_d and P_{fa}. Upon closer examination, however, they are quite different. While P_d and P_{fa} summarize how much of the truth causes an alarm, *PPV* and *NPV* summarize how many alarms turn out to be true.

PPV and *NPV* can also be easily calculated from the 2 × 2 confusion matrix. From an operator's perspective, our notional air cargo screening system exhibits a conflicting performance:

$$NPV = \frac{TN}{TN + FN} = \frac{9{,}500}{9{,}500 + 10} \approx 1 \text{ (compared to 1 for a perfect system)} \quad (3)$$

$$PPV = \frac{TP}{TP + FP} = \frac{90}{90 + 500} = 0.15 \text{ (compared to 1 for a perfect system)} \quad (4)$$

Equation 3 shows that, of all items that did *not* cause an alarm ($TN + FN = 9{,}500 + 10 = 9{,}510$), a large subset ($TN = 9{,}500$) correctly turned out to *not* contain a true threat. These counts resulted in $NPV \approx 1$, approximately equal to the 1 value that would be exhibited by a perfect system.[4] In the absence of an alarm, the operator could rest assured that a threat was highly unlikely. However, Equation 4 shows that, of all items that did indeed cause an alarm ($TP + FP = 90 + 500 = 590$), only a small subset ($TP = 90$) turned out to contain a true threat (i.e., were not false alarms). These counts unfortunately led to $PPV = 0.15$, much lower than the 1 value that would be exhibited by a perfect system. When an alarm was declared, the operator could not trust that a threat was present, since the system cried wolf so often.

[3] For P_d and P_{fa} values from equations (1) and (2), statistical tests can determine whether the system's value is significantly different from the perfect value and if it is different from the capability requirement (Fleiss et al. 2013).

[4] For *NPV* and *PPV* values from equations (3) and (4), statistical tests can determine whether the system's value is significantly different from the perfect value and if it is different from the capability requirement (Fleiss et al. 2013.

Reconciling Developers with Operators: P_d and P_{fa} versus *PPV* and *NPV*

The discrepancy between *PPV* and *NPV* versus P_d and P_{fa} reflects the discrepancy between operators' and developers' views of the threat detection system. Developers are often primarily interested in how much of the truth correctly cause alarms—concepts quantified by P_d and P_{fa}. In contrast, operators are often primarily concerned with how many alarms turn out to be true—concepts quantified by *PPV* and *NPV*. As shown in Figure 2, the very same system that exhibits excellent values for P_d, P_{fa}, and *NPV* can also exhibit poor values for *PPV*.

Poor *PPV* values can be expected for DoD and DHS threat detection systems. Such performance is often merely a reflection of the low *Prev* of true threats among true clutter that commonly occurs in defense and homeland security scenarios.[5] *Prev* describes the fraction of all items that contain a true threat, including those that did and did not cause an alarm. In the case of our notional air cargo screening system, *Prev* is very low:

$$Prev = \frac{TP + FN}{TP + FN + FP + TN} = \frac{90 + 10}{90 + 10 + 500 + 9{,}500} = 0.01. \quad (5)$$

Equation 5 shows that, of all items ($TP + FN + FP + TN = 90 + 10 + 500 + 9{,}500 = 10{,}100$), only a small subset ($TP + FN = 90 + 10 = 100$) contained a true threat, leading to $Prev = 0.01$. When true threats are rare, most alarms turn out to be false, even for an otherwise strong threat detection system, leading to a low value for *PPV*. In fact, to achieve a high value of *PPV* when *Prev* is extremely low, a threat detection system must exhibit so few *FP*s (false alarms) as to make P_{fa} approximately zero.

Recognizing this phenomenon, PMs should not necessarily dismiss a threat detection system simply because it exhibits a poor *PPV*, provided that it also exhibits an excellent P_d and P_{fa}. Instead, PMs can estimate *Prev* to help determine how to guide such a system through development. *Prev* does not depend on the threat detection system and can, in fact, be calculated in the absence of the system. Knowledge of ground truth (i.e., which items contain a true threat) is all that is needed to calculate *Prev* (Scheaffer and McClave 1995).

Of course, ground truth is not known *a priori* in an operational setting. However, it may be possible for PMs to use historical data or intelligence tips to roughly estimate whether *Prev* is likely to be particularly low in operation. A *Prev* that is estimated to be particularly low can cue the PM to anticipate discrepancies in P_d and P_{fa} versus *PPV*, forecasting the inevitable discrepancy between the developers' versus operators' views early in the system's

[5] Conversely, when *Prev* is *high*, threat detection systems often exhibit poor values for *NPV*, even while exhibiting excellent values for P_d, P_{fa}, and *PPV*. Such cases are not discussed here, since fewer scenarios in DoD and DHS involve a *high* prevalence of threat among clutter.

development, while there are still time and opportunity to make adjustments. At that point, the PM can identify concepts of operations in which the system can still provide value to the operator for his or her mission. A tiered system may provide one such opportunity.

References

Altman, D. G., and J. M. Bland. 1994. "Diagnostic Tests 2: Predictive Values." *British Medical Journal* 309, no. 6947: 102. https://doi.org/10.1136/bmj.309.6947.102.

Bliss, J. P., R. D. Gilson, and J. E. Deaton. 1995. "Human Probability Matching Behavior in Response to Alarms of Varying Reliability." *Ergonomics* 38, no. 11: 2300–2312. https://doi.org/10.1080/00140139508925269.

Cushman, J. H. 1987. "Making Arms Fighting Men Can Use." *New York Times*. June 21. http://www.nytimes.com/1987/06/21/business/making-arms-fighting-men-can-use.html.

Fleiss, J. L., B. Levin, and M. C. Paik. 2013. *Statistical Methods for Rates and Proportions* 3rd ed. Hoboken, NJ: John Wiley.

Getty, D. J., J. A. Swets, R. M. Pickett, and D. Gonthier. 1995. "System Operator Response to Warnings of Danger: A Laboratory Investigation of the Effects of the Predictive Value of a Warning on Human Response Time." *Journal of Experimental Psychology: Applied* 1, no. 1: 19–33. https://doi.org/10.1037/1076-898X.1.1.19.

Oldham, J. 2006. "Outages Highlight Internal FAA Rift." *Los Angeles Times*. October 3. http://articles.latimes.com/2006/oct/03/local/me-faa3.

Parasuraman, R. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* 39, no. 2: 230–253. https://doi.org/10.1518/001872097778543886.

Scheaffer, R. L., and J. T. McClave. 1995. "Conditional Probability and Independence: Narrowing the Table." In *Probability and Statistics for Engineers*, 85–92. 4th ed. Belmont, CA: Duxbury Press.

Stuart, R. 1987. "U.S. Cites Amtrak for Not Conducting Drug Tests."' *The New York Times*. January 8. http://www.nytimes.com/1987/01/08/us/us-cites-amtrak-for-not-conducting-drug-tests.html.

Urkowitz, H. 1967. "Energy Detection of Unknown Deterministic Signals." *Proceedings of the IEEE* 55, no. 4:523–531. MV{.

Shelley Cazares, a Research Staff Member in the Science and Technology Division of IDA's Systems and Analyses Center, holds a doctorate in engineering science from the University of Oxford.