How Certification Systems Fail:

Lessons from the Ware Report

Steven J. Murdoch, Mike Bond, and Ross Anderson | University of Cambridge

The 1970 *Security Controls for Computer Systems* report, which helped shape computer systems' standard evaluation criteria, can shed light on current certification systems' shortcomings.

Most security certification standards in the payments industry are based on *Security Controls for Computer Systems*, a 1970 US Department of Defense task force report. Commonly called the "Ware Report" (for the chair of the task force, Willis H. Ware), it focused on protecting classified information in multiaccess, resource-sharing computer systems that were increasingly used by both the government and its defense contractors.[1] The report included recommendations for security functionality, such as systems should have to process classified information safely, as well as proposed certification procedures for verifying whether a system meets various criteria. These certification procedures ultimately formed the basis for the *Trusted Computer System Evaluation Criteria* (TCSEC) report, whose requirements and assessment criteria are listed in 5200.28-STD, colloquially known as the "Orange Book."[2] The "Rainbow Series" augmented this latter publication, expanding and clarifying various aspects of it.

Although eventually superseded, TCSEC was highly influential in the development of two widely used certification standards in the payments industry: FIPS-140 and Common Criteria. FIPS-140 defines security requirements for cryptographic modules (both software and hardware) and includes specification of cryptographic functionality and tamper-resistance measures. However, for software-based cryptographic modules, FIPS-140 requires that the operating system be certified to the TCSEC standard. Common Criteria replaced TCSEC (along with a few other standards) but is far broader. Rather than ensuring only the confidentiality of classified information, Common Criteria evaluates a variety of systems, including identification schemes for trash cans (for when households are billed for sanitation services based on how full their cans are) and devices to stop drunk drivers from starting their cars.

In the payments industry, FIPS-140 and Common Criteria are used to evaluate smart cards and hardware security modules (HSMs—add-on cards to computers that store cryptographic keys and restrict the operations performed on them according to a security policy). However, Common Criteria also evaluates ATMs and point-of-sale terminals as well as equipment less specific to the payments industry, such as firewalls and intrusion detection systems.

Complying with these standards is onerous, and the certification process is both expensive and time-consuming. However, security vulnerabilities are regularly discovered in all these systems, some of which are easy to exploit. How were these flaws missed? Was it a failure of the evaluation or a failure in the evaluation

Copublished by the IEEE Computer and Reliability Societies

1540-7993/12/$31.00 © 2012 IEEE

scheme? We can answer some of these questions by looking back at the report that originated TCSEC and its descendants.

Who Performs the Evaluation?

The Ware Report states[1]

> Any computer system used to process classified information shall be subjected to inspection and test by expert technical personnel acting for the Responsible Authority. The extent and duration of the inspections and tests shall be at the discretion of the Responsible Authority. The inspections and tests shall be conducted to determine the degree to which the system conforms to the requirements here recommended, any derivative regulations, and other applicable regulations.

The Ware Report recommends that the evaluation be performed on behalf of a responsible authority (the "head of the department or agency responsible for the proper operation of the secured computer system"[1]). Indeed, this was how TCSEC was implemented: as the US government agency responsible for protecting classified information, the US National Security Agency (NSA) carried out product evaluation.

In contrast, FIPS-140 and Common Criteria evaluation are performed by a commercial testing laboratory that the product vendor selects. This introduces a clear conflict of interest—vendors will want to select a lab that gives their product an easy ride (for example, by asking fewer questions or doing the evaluation more quickly). The designers of Common Criteria and FIPS-140 recognized this conflict, and in both schemes, it is the relevant government body that actually grants certification (though mainly on the basis of the testing lab's report).

In addition, the relevant government body licenses the evaluation labs; labs that consistently fail to achieve adequate standards risk having their right to perform evaluations revoked. This threat is intended to prevent a race to the bottom in evaluation lab standards, but it's far from perfect: we know of no case in which a lab's license was revoked, and labs that maintain high standards complain about losing business.

The situation is even worse when participants actively try to subvert the certification process. We've seen this in the case of PIN entry device (PED) evaluation. PEDs are used at point-of-sale terminals and frequently incorporate a smart-card reader. Currently, more than 1.34 billion smart cards are in circulation, and PEDs are becoming increasingly common.

PEDs mustn't allow the addition of a device that can capture the customer's PIN, because with the PIN and card details, attackers can create a duplicate magnetic stripe card for use in ATMs that haven't been upgraded with smart-card readers. The Common Criteria specification for PEDs says the following:[3]

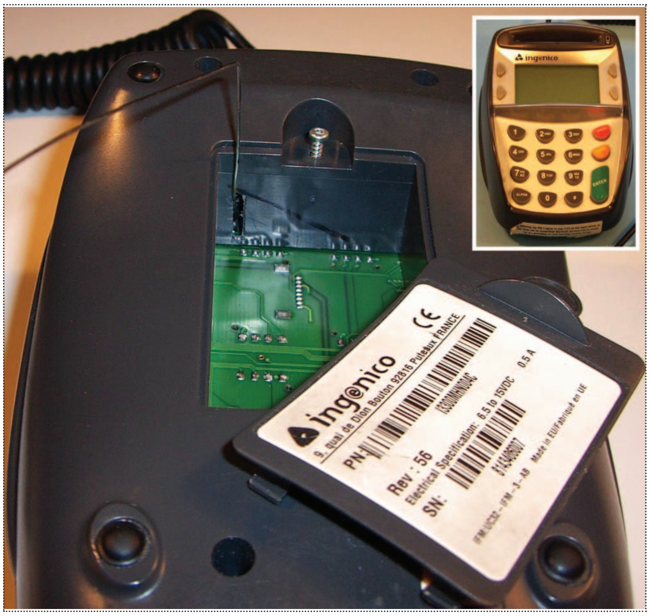

Figure 1. Despite being advertised as having passed Common Criteria evaluation, some products on the market, such as the Ingenico i3300 PIN entry device, allow insertion of PIN-tapping devices.

> The [Target of Evaluation Security Functions] shall resist physical attacks based on addition of any PIN tapping device to the PIN Entry Device and Card Reader by … providing the capability to detect such attacks with a high probability [or] automatically responding such that the [Target of Evaluation Security Policy] is not violated.

In Common Criteria jargon, the specification requires that PED PIN-tapping devices be detected. Yet, we've proven that numerous PEDs on the market have flaws that allow the capture of PINs and card details. One such PED is the Ingenico i3300 (see Figure 1), which even comes equipped with a rear compartment in which attackers can store a tapping device. All they need to do is to cut a small hole in the case and hook a paper clip onto a communication line over which unencrypted PINs and card details are sent. Criminals have performed similar attacks on this terminal.[4]

Ingenico i3300 is one of several PEDs that are advertised as having passed Common Criteria evaluation, yet it's trivial to tamper with. What went so badly wrong? When we investigated, we couldn't determine which lab evaluated the PED and were refused a copy of the certification report—both of which should be publicly available information. We later found that the evaluation wasn't performed properly: a licensed testing lab evaluated the PED but didn't involve a government body permitted to issue Common Criteria certifications.

Skipping this critical final step removes the pressure on testing labs to do a good job. How can their license be revoked when we can't find out their name? The UK payments industry's representative body, the Association for Payment Clearing Services (APACS), which operates this pseudo–Common Criteria scheme, terms devices they approve as Common Criteria "evaluated" as opposed to "certified," but this doesn't stop vendors from claiming that devices are Common Criteria certified. The UK government body responsible for Common Criteria—Government Communications Headquarters (GCHQ)—appears uninterested in protecting the Common Criteria brand by preventing such behavior.

"Ideally, certification efforts concentrate on those components on which most assurance is necessary, but economic pressures lead to the most expedient certification."

Composing Evaluations

Certification is expensive and frequently performed only on part of a system. Ideally, certification efforts concentrate on those components on which most assurance is necessary, but economic pressures lead to the most expedient certification. The Ware Report cautions about such combinations of certified and uncertified components:[1]

> It is not certain at the present time that tests can adequately establish the integrity of boundaries, thus permitting inclusion of an uncertified portion in a system. In general, the more highly classified and sensitive the information in a system, the more carefully one should consider the risks before permitting an uncertified portion to operate in the overall system.

One device that was found to be insecure despite being certified was the IBM 4758 HSM. It achieved FIPS-140 level 4 certification (the highest level possible) following stringent evaluation of its tamper-resistance measures and cryptographic functionality. Attackers wishing to extract keys from an IBM 4758 through physical tampering would need to deal with multiple layers of tamper-detecting mesh, epoxy potting, temperature sensors, and x-ray detectors. However, the evaluation didn't include the software loaded on the device—the IBM Common Cryptographic Architecture (CCA).

The CCA was designed to ensure that no single person could initiate a procedure that would compromise the security of the most sensitive keys by requiring that the keys it generates are split into two parts and given to two different people. Encryption and authentication in the payments industry normally use 3DES with two 56-bit keys (that is, a 112-bit key). 3DES is secure enough for most purposes, but for backward compatibility, the IBM 4758 device also supports single DES (56-bit keys) that can be broken easily by brute force.

The CCA wisely restricts how single DES can be used. In particular, although it's permitted to extract a 112-bit key encrypted under 3DES with another 112-bit key, it should be impossible to extract a 112-bit key encrypted under single DES (that is, with a 56-bit key). So, moving keys between HSMs should be possible, but no individual should be able to establish a key's cleartext value. However, with some trickery, completely circumventing the tamper protection and extracting a 112-bit key is possible.

To understand the flaw that allows this, we need to know how 3DES with two keys is built on single DES. 3DES 112-bit keys have a left half and a right half (each 56 bits). Encryption proceeds by first performing a single DES encryption under the left key, then a single DES decryption under the right key, and finally a single DES encryption under the left key. If the key's left and right halves are different, this construction is stronger than single DES. But, if they're the same, the middle decryption stage cancels out the first encryption stage, resulting in a single DES encryption.

This construction gives 3DES the desired backward compatibility with single DES but also opens up a vulnerability. Attackers can get the IBM 4758 device to generate two 56-bit keys and discover their cleartext value by brute force. (This took two days in 2001 when this vulnerability was discovered;[5] today, it would take a lot less time.)

Because the CCA doesn't bind the two halves of keys, attackers can create a 3DES 112-bit key whose left half is one of the known 56-bit keys and the right half is the other known 56-bit key. They've now created a 112-bit 3DES key with different left and right halves that the device will let them use to encrypt other 112-bit 3DES keys. Because attackers know the value of both of the 56-bit single DES keys they created, they know the value of the 112-bit 3DES key. They can simply decrypt all the keys they extracted in encrypted form and break the system's security.

This attack is subtle but has been proven possible,[5] despite the effort put into creating a secure device and performing the FIPS-140 evaluation. The flaw exploited wasn't spotted because the evaluation dealt only with the hardware and core software, not the CCA, which enforces the security properties that banks care about.

Is the Evaluation Appropriate?

IBM 4758 failed because the FIPS-140–evaluated hardware used uncertified CCA software, and the result was insecure. This is a specific example of a more general problem—determining whether the certified system's environment is sufficient to achieve the overall security goals. To incorporate this into the evaluation process, the Ware Report proposed three types of certification:[1]

> ***Design Certification***. A series of tests and inspections that establish that the safeguards designed into the hardware and software of the system are operative, function as intended, and collectively constitute acceptable controls for safeguarding classified information. Production models of a given design need be tested only to verify that all safeguards are present and properly functioning. …
>
> ***Installation Certification***. A series of tests and inspections performed according to specifications established during the design certification phase to insure that the required set of security safeguards (hardware, software, and procedural) are in fact present and operational in the installed equipment, and on all communication links that will carry classified information to remote terminals or other computers. This certification must also examine the operational procedures and administrative structure of the organization that controls the equipment, and must establish that the procedural and administrative environment supplements and complements hardware and software safeguards, and that physical safeguards are appropriate. …
>
> ***Recertification***. Some level of recertification must be accomplished periodically, as indicated by operational circumstances. …

TCSEC, FIPS-140, and Common Criteria are design certifications. They make assertions about whether a product can fulfill some security property (in the case of the Ware Report and TCSEC, safeguarding classified information) but can't make general claims about security. Even if the security properties evaluated match those the system needs to maintain, without installation certification, it's difficult to say the system fulfills the security properties in the real world.

In the payments industry, card providers such as Visa and MasterCard operate their own certification schemes incorporating some of the installation certification's tests. However, the processes and results aren't made public (unlike properly performed FIPS-140 and Common Criteria certifications).

The Ware Report doesn't discuss certification reports being made public, but we can see the reason by returning to the definition of the responsible authority that manages the certification process: "head of the department or agency responsible for the proper operation of the secured computer system."[1] Note that "department" and "agency" are singular—it's implied that only one department or agency is responsible for the computer system. This is appropriate for classified information processing in the US, where there is one agency (the NSA) with overall responsibility for securing classified information. In contrast, the responsibility for securing payment systems is diffuse—including banks, card schemes, hardware/software vendors, and customers. When something goes wrong and fraud occurs, one of these parties must take the blame, and frequently, it's the customer.

In cases we've dealt with, customers disputing transactions in their account are sometimes accused of negligence (for example, not adequately protecting their card or PIN); on other occasions, banks have even accused customers of deliberately committing fraud by making false statements. Banks ask courts and adjudicators to rely on system certification as evidence that the banks' conclusions are correct yet, frequently, don't disclose the certification reports. Without access to the reports, courts and adjudicators can't identify the certification process's limitations and customers can't effectively obtain expert help in interpreting bank-submitted evidence. Requiring public certification reports (as for Common Criteria and FIPS-140) helps correct this problem. Whereas making certification reports available only to a system's owner might be acceptable in the situations envisaged by the Ware Report, the complicated multi-stakeholder payments industry environment requires a different approach.

Lessons Learned

Despite intending to secure classified information in military environments, the Ware Report has much to teach today's certification process designers and implementers. The report shows that many of the challenges facing current certification schemes have been known for more than 40 years. The security research community has made progress in resolving some of these, but some of the lessons have been lost along the way.

Researchers continue to discover security problems that result from the composition of certified and uncertified components. The Ware Report stated that tests might be unable to establish the integrity of security boundaries. Although there have been advances in understanding the composition of certain classes of components (for example, cryptographic protocols), no general technique exists for reasoning about systems built from components of differing trustworthiness. The Ware Report's cautions—against permitting uncertified components from operating in systems processing sensitive information—remain prudent.

The Ware Report's concept of installation certification also remains valuable today. Often, the task of establishing whether a certified product operates appropriately is an afterthought and carried out with a far lower level of rigor than that of the product's design certification. The payments industry needs a greater appreciation that merely using certified products isn't sufficient to maintain a secure system.

The question of how to recertify products also deserves revisiting. The Ware Report said this should be performed "as indicated by operational circumstances,"[1] and it's likely that appropriate recertification procedures will vary depending on the product type being certified. It's clear that PED certification practices could be improved—certifications do expire, but by that time, the product will likely have been discontinued. Triggering recertification when an advance in criminal capability is discovered would also be prudent.

Although there were good economic reasons for moving from the single certification body of TCSEC to the marketplace of commercial certification labs for FIPS-140 and Common Criteria, this decision should be reevaluated continually. In cases in which parties other than the system owner are asked to rely on the quality of certification, perhaps manufacturers shouldn't be given a free hand in deciding the certification type and level or choosing the laboratory that performs the test.

Revisiting the lessons presented in the Ware Report can help us improve the quality of certification, but we shouldn't expect certification to be a silver bullet:[1]

> The security problem of specific computer systems must, at this point in time, be solved on a case-by-case basis, employing the best judgment of a team consisting of system programmers, technical hardware and communications specialists, and security experts.

Ware's summary statement in 1970 remains as true today as it was then. ■

> In the early 1960s, Peter Naur devised an elegant runtime system for his Algol 60 compiler in which recursion was implemented by a push-down stack that contained both data and program. This turned out to be possibly the worst thing ever to happen to the computer security field.
>
> —Earl Boebert

References

1. W.H. Ware, *Security Controls for Computer Systems: Report of Defense Science Board Task Force on Computer Security*, report R-609-1, RAND, Jan. 1970.
2. *Trusted Computer System Evaluation Criteria*, US Dept. of Defense, Nat'l Computer Security Center, report 5200.28-STD, Dec. 1985.
3. *APACS PIN Entry Device Protection Profile* (version 1.37), Assoc. for Payment Clearing Services (APACS), July 2003.
4. S. Drimer, S.J. Murdoch, and R. Anderson, "Thinking Inside the Box: System-Level Failures of Tamper Proofing," *IEEE Symp. Security and Privacy*, IEEE CS, 2008, pp. 281–295.
5. R. Clayton and M. Bond, "Experience Using a Low-Cost FPGA Design to Crack DES Keys," *Proc. Workshop on Cryptographic Hardware and Embedded Systems* (CHES 02), LNCS 2523, Springer, 2002, pp. 877–883.

Steven J. Murdoch is a researcher at the University of Cambridge Computer Laboratory and a Fellow of Christ's College. He conducts research on bank payment system security and helps fraud victims. He has also worked extensively on the Tor Project, researching and designing safe communication systems for people living and working in repressive regimes and developing technology that circumvents censorship. Murdoch has a PhD in computer security from the University of Cambridge. Contact him at Steven.Murdoch@cl.cam.ac.uk.

Mike Bond is a visiting industrial fellow at the University of Cambridge. His research interests include EMV, banking security, tamper-resistant device security, and cheating in computer games. Bond has a PhD in computer security from the University of Cambridge. Contact him at Mike.Bond@cl.cam.ac.uk.

Ross Anderson is professor of security engineering at the University of Cambridge. He was one of the founders of security economics and was a pioneer of peer-to-peer systems, hardware tamper-resistance, copyright marking, and API security. Anderson has a PhD in computer security from the University of Cambridge. He chairs the Foundation for Information Policy Research; is a Fellow of the Royal Society, the Royal Academy of Engineering, the IET, the IMA, and the Institute of Physics; and wrote the definitive textbook *Security Engineering—A Guide to Building Dependable Distributed Systems* (Wiley, 2008). Contact him via www.ross-anderson.com.