United States Government Accountability Office

GAO

Report to Congressional Committees

September 2008

COMBATING NUCLEAR SMUGGLING

DHS's Phase 3 Test Report on Advanced Portal Monitors Does Not Fully Disclose the Limitations of the Test Results

GAO-08-979

September 2008

COMBATING NUCLEAR SMUGGLING

DHS's Phase 3 Test Report on Advanced Portal Monitors Does Not Fully Disclose the Limitations of the Test Results

GAO Accountability · Integrity · Reliability

Highlights

Highlights of GAO-08-979, a report to congressional committees

Why GAO Did This Study

The Department of Homeland Security's (DHS) Domestic Nuclear Detection Office (DNDO) is responsible for addressing the threat of nuclear smuggling. Radiation detection portal monitors are part of the U.S. defense against such threats. In 2007, Congress required that funds for new advanced spectroscopic portal (ASP) monitors could not be spent until the Secretary of DHS certified that these machines represented a significant increase in operational effectiveness over currently deployed portal monitors. In addition to other tests, DNDO conducted the Phase 3 tests on ASPs to identify areas in which the ASPs needed improvement. GAO was asked to assess (1) the degree to which the Phase 3 test report accurately depicts the test results and (2) the appropriateness of using the Phase 3 test results to determine whether ASPs represent a significant improvement over current radiation detection equipment. GAO also agreed to provide its observations on special tests conducted by Sandia National Laboratories (SNL).

What GAO Recommends

GAO's recommendations include proposing that the Secretary of DHS revise the Phase 3 report to better disclose test results and limitations if it is to be used in any certification decision for ASP acquisition. DHS disagreed with two of GAO's recommendations but agreed to take action on a third. GAO continues to believe that all of its recommendations need to be implemented.

To view the full product, including the scope and methodology, click on GAO-08-979. For more information, contact Gene Aloise at 202-512-3841 or aloisee@gao.gov.

What GAO Found

Because the limitations of the Phase 3 test results are not appropriately stated in the Phase 3 test report, the report does not accurately depict the results from the tests and could potentially be misleading. In the Phase 3 tests, DNDO performed a limited number of test runs. Because of this, the test results provide little information about the actual performance capabilities of the ASPs. The report often presents each test result as a single value; but considering the limited number of test runs, the results would be more appropriately stated as a range of potential values. For example, the report narrative states in one instance that an ASP could identify a source material during a test 50 percent of the time. However, the narrative does not disclose that, given the limited number of test runs, DNDO can only estimate that the ASP would correctly identify the source from about 15 percent to about 85 percent of the time—a result that lacks the precision implied by DNDO's narrative. DNDO's reporting of the test results in this manner makes them appear more conclusive and precise than they really are. The purpose of the Phase 3 tests was to conduct a limited number of test runs in order to identify areas in which the ASP software needed improvement. While aspects of the Phase 3 report address this purpose, the preponderance of the report goes beyond the test's original purpose and makes comparisons of the performance of the ASPs with one another or with currently deployed portal monitors.

In GAO's view, it is not appropriate to use the Phase 3 test report in determining whether the ASPs represent a significant improvement over currently deployed radiation equipment because the limited number of test runs do not support many of the comparisons of ASP performance made in the Phase 3 report. As the report shows, if an ASP can identify a source material every time during a test, but the test is run only five times, the only thing that can be inferred with a high level of statistical confidence is that the probability of identification is no less than about 60 percent. Although DNDO states in the Phase 3 test report that the results will be relevant to the Secretary's certification that the ASPs represent a significant increase in operational effectiveness, it does not clarify in what ways the results will be relevant. Furthermore, DNDO offers no explanation as to why it changed its view from the Phase 3 test plan, which states that these tests will not be used to support a certification decision.

The goal of SNL's special tests was, among other things, to identify potential vulnerabilities in the ASPs by using different test scenarios from those that DNDO planned to use in other ASP tests. SNL concluded in its test report that the ASPs' software and hardware can be improved and that rigor could be added to DNDO's testing methods. Furthermore, the report acknowledges that (1) a specific objective of the testing at the Nevada Test Site was to refine and improve the ASP's performance and (2) the special tests were never intended to demonstrate conformity of the ASPs with specific performance requirements. In GAO's view, these statements appear to accurately describe the purpose, limitations, and results of the special tests.

Contents

Abbreviations

ASP	advanced spectroscopic portal
CBP	Customs and Border Protection
DHS	Department of Homeland Security
DNDO	Domestic Nuclear Detection Office
PVT	polyvinyl toluene
RIID	radioactive isotope identification device
SNL	Sandia National Laboratories

GAO
Accountability * Integrity * Reliability

United States Government Accountability Office
Washington, DC 20548

September 30, 2008

The Honorable John D. Dingell
Chairman
The Honorable Joe Barton
Ranking Member
Committee on Energy and Commerce
House of Representatives

The Honorable Bart Stupak
Chairman
The Honorable John Shimkus
Ranking Member
Subcommittee on Oversight and Investigations
Committee on Energy and Commerce
House of Representatives

Preventing a nuclear weapon or radiological dispersal device (a "dirty bomb") from being smuggled into the United States is a key national security priority. The Department of Homeland Security (DHS), through its Domestic Nuclear Detection Office (DNDO), has lead responsibility for conducting the research, development, testing, and evaluation of equipment that can be used to detect smuggled nuclear or radiological materials. U.S. Customs and Border Protection (CBP) is responsible for screening cargo as it enters the nation at our borders, including operating radiation detection equipment to intercept dangerous nuclear and radiological materials.

Much of DNDO's work on radiation detection equipment has focused on the development and use of radiation detection portal monitors that can screen vehicles, people, and cargo entering the United States. In the case of cargo, these portal monitors typically include two detector panels that form sidewalls of a portal through which trailer trucks carrying cargo containers must pass. Currently, CBP employs portal monitors made of polyvinyl toluene (a plastic), known as PVTs, which can detect the presence of radiation but cannot identify the specific radiological material generating the radiation. As a result, PVTs cannot distinguish between benign, naturally occurring radiological materials such as ceramic tile, and dangerous materials such as highly enriched uranium. Therefore, if a PVT detects the presence of radiation in a shipping container during a primary inspection, CBP conducts a second inspection with another PVT and uses a handheld radioactive isotope identification device (RIID) to identify the type of source material emitting radiation. However, RIIDs use detectors

that are relatively small and, as a result, are limited in their ability to correctly identify radiological and nuclear source materials, so CBP officials sometimes must consult with scientists at CBP's Laboratories and Scientific Services or physically search a container to identify the radiation source. Nonetheless, CBP officials have stated that the current system of radiation detection equipment and standard operating procedures provide the best possible radiological and nuclear screening coverage available with current technology and that it does not have a significant adverse impact on the flow of commerce.

In 2005, in an effort to overcome the technical limitations of PVTs and RIIDs, DNDO sponsored research, development, and testing intended to produce Advanced Spectroscopic Portal (ASP) monitors, which are designed to both detect the radiological or nuclear source material and identify the specific type of source material. According to DNDO, ASPs will reduce the number of false positives or nuisance alarms—instances in which a portal monitor detects the presence of radiation and signals an alarm, but the source material turns out to be benign.

In July 2006, DNDO awarded contracts worth potentially $1.2 billion to three vendors to further develop and produce ASPs over five years. Shortly thereafter, Congress required in DHS's fiscal year 2007 appropriation that the Secretary of Homeland Security certify that ASPs represent "a significant increase in operational effectiveness" before DNDO could order full-scale production. Congress enacted a similar requirement in DHS's fiscal year 2008 appropriation.

In early 2007, DNDO conducted formal tests on ASPs in two phases at the Nevada Test Site. The first, called Phase 1, was designed to assess the ASP's performance capabilities in order to support a full-scale production decision with a high degree of statistical confidence. DNDO told us on multiple occasions and in a written response that only data collected during Phase 1 would be included in the final report presented to the Secretary to request ASP certification. According to DNDO, the second group of tests, called Phase 3, provided data for algorithm (software) improvement that targeted specific and known areas in need of work and data to aid in the development of secondary screening operations and procedures.[1] For example, the Phase 3 tests attempt to determine, among other things, how ASP performance is affected by the presence of various

[1]According to DNDO, Phase 2 was its completion of the test report for the Phase 1 tests.

substances—known as shielding materials—that partially absorb or alter the radiation emitted by the source material in a container. During the Phase 3 tests, DNDO also tested the ASPs to determine how their performance changes when the container moves through the portal monitor at different speeds.

The size of the samples used in the Phase 1 and Phase 3 tests are important in determining the confidence one can place in the test results.[2] Larger sample sizes, such as the 15 to 60 test runs performed during the Phase 1 tests, usually allow a more precise interpretation of results, i.e., estimates of ASP performance may fall within a fairly narrow range of values. Conversely, estimates drawn from small sample sizes, such as the 1 to 10 test runs done for Phase 3, normally have much less precision associated with them—thus the range of potential values may be quite wide. According to DNDO's Phase 3 test plan, the Phase 3 testing consisted of small sample sizes to allow more time to test a wide range of source materials in order to make improvements to the ASPs. However, the small samples associated with the Phase 3 tests would make it difficult to use the test results as a reliable indicator of the ASPs' performance capabilities or for comparisons of performance among various detection systems. In contrast, the Phase 1 tests involved larger sample sizes so that DNDO could assess the performance capabilities of the ASPs with a higher degree of statistical precision. Because of the small sample sizes, the Phase 3 test plan stated that Phase 3 testing would not be used to support a full-scale production decision.

In September 2007, we testified that, in our view, DNDO's Phase 1 tests did not represent an objective or rigorous assessment of the ASPs. More specifically, we stated that DNDO used biased test methods that enhanced the apparent performance of the ASPs and did not identify the limitations

[2]Equipment testing involves repeating a single test multiple times to estimate how often a device performs its function correctly. The set of tests is referred to as a sample. Each test in the sample is considered a random trial and therefore the estimates derived from the sample are subject to random variations; in other words, if the series of tests were repeated then each sample could yield different estimates. Because of these random variations, estimates from samples are often presented as a range of possible values called the 95 percent confidence interval. This is the range of values that will contain the true probability of performance in 95 percent of the samples that we might select. In general, when the sample size (number of tests) is larger, the range of possible values is smaller, allowing more precise estimates of the likely performance of the machine.

of the ASPs' detection capabilities.[3] During that hearing, DNDO's Director changed DNDO's position and stated that the Secretary could use the Phase 3 tests in combination with other test results when deciding whether to certify the ASPs.

In its report on the Phase 3 tests, DNDO states that

- The ASPs were as good as or better than the PVTs at detecting the presence of radiological source materials at low levels of radiological activity.
- The performance of the ASPs from each of the three vendors was statistically indistinguishable with few exceptions, for each category of source material (i.e., medical sources, industrial sources, and special nuclear material).
- The performance of the RIIDs was poor compared to the performance of the ASPs in identifying the specific radiological or nuclear source material inside a container.

At the same time the Phase 1 and Phase 3 tests were ongoing, Sandia National Laboratories (SNL), at the request of DNDO, conducted a series of tests that would go beyond those covered in the Phase 1 and Phase 3 tests. The goal of these tests, called special tests, was, among other things, to identify potential vulnerabilities in the ASPs by using source materials and test scenarios different from those that DNDO planned to use in either the Phase 1 or the Phase 3 tests. The tests were "blind" in that neither the ASP vendors nor the ASP test operators knew what was in the containers being tested. The review was also to focus on vulnerabilities in the test processes DNDO described in its Phase 1 and Phase 3 test plans.

Given DNDO's change in how it believes the Phase 3 test results may be applied, the significant costs of the ASPs, and the importance of protecting our borders from nuclear smuggling you asked us to assess (1) the degree to which the Phase 3 test report accurately depicts the test results and (2) the appropriateness of using the Phase 3 test results to determine whether ASPs represent a significant improvement over current radiation detection

[3]GAO, *Combating Nuclear Smuggling, Additional Actions Needed to Ensure Adequate Testing of Next Generation Radiation Detection Equipment*, GAO-07-1247T (Washington, D.C.: Sept. 18, 2007).

equipment. We also agreed to provide our observations on the special tests conducted by SNL.

To perform our work, we reviewed the Phase 3 test report and SNL's special tests report. We met with key officials from the National Institutes of Standards and Technology who were responsible for designing part of the Phase 3 tests and analyzing their results. We also relied on documents and other evidence gathered during our previous review of ASP testing at the Nevada Test Site. We conducted this performance audit from February 2008 to August 2008 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Results in Brief

Because the limitations of the Phase 3 test results are not appropriately stated in the Phase 3 test report, the report does not accurately depict the results from the tests and could be misleading. In the Phase 3 tests, DNDO performed a limited number of test runs. Because of this, the test results provide little information about the actual performance capabilities of the ASPs. The report narrative often presents test results using a single value or percentage. Considering the limited number of test runs, the results would be more appropriately stated as a range of potential values. This important limitation is apparent only by reviewing more technical data elsewhere in the report. For example, the report narrative states in one instance that an ASP could identify a source material during a specific test 50 percent of the time. However, the narrative does not disclose that, given the limited number of test runs, DNDO can only estimate that the ASP would be able to correctly identify the source from about 15 percent to about 85 percent of the time—a result that lacks the precision implied by DNDO's narrative. DNDO's reporting of the test results in this manner makes the results appear more conclusive and precise than they really are. For example, the executive summary states that the ASPs "demonstrated detection limits equal to or better than those of any of the PVT systems," but fails to mention that the test used only a single source material and that the results would not necessarily be the same for other radiological sources. In fact, the Phase 3 results showed that the PVTs could detect some source materials much better than the ASPs. The purpose of the Phase 3 tests was to identify areas in which the ASP software needed improvement. While aspects of the Phase 3 report address this purpose, the preponderance of the report goes beyond the test's original purpose

and makes comparisons of the performance of the ASPs with each other or with PVTs and RIIDs.

In our view, it is not appropriate to use the Phase 3 test report in determining whether the ASPs represent a significant improvement over currently deployed radiation equipment because the limited number of test runs does not support many of the comparisons of ASP performance made in the Phase 3 report. As the report shows, if an ASP can identify a source material every time during a test, but the test is run only five times, the only thing that can be inferred at the 9 percent confidence level is that the true probability of identification is no less than approximately 60 percent. Although DNDO states in the Phase 3 test report that the test results will be relevant to the Secretary's certification that the ASPs represent a significant increase in operational effectiveness, the report does not clarify in what ways the results will be relevant. Furthermore, DNDO offers no explanation as to why it changed its view from the Phase 3 test plan which states that the Phase 3 tests will not be used to support a certification decision.

Regarding SNL's special tests, SNL concluded in its test report that the ASPs' software and hardware can be improved in some areas and that rigor could be added to DNDO's testing methods. Furthermore, the report acknowledges that (1) a specific objective of the testing at the Nevada Test Site was to refine and improve the ASP software performance and (2) the special tests "were never intended to demonstrate conformity of the [ASP] systems against specific performance requirements." In our view, these statements appear to accurately describe the purpose, limitations, and results of the special tests. In addition, the report concluded that upon reviewing data from DNDO's previous ASP tests, the reported results were consistent with the data that DNDO collected and that as a result, "DNDO's ASP system assessment was not biased." It is important to note, however, SNL's report also concluded that the "ASP system assessment [in 2007] was not biased" and that SNL "observed no data suggesting that the ASP system performance was inappropriately manipulated." In making this statement, SNL is referring to the data derived from ASP tests. In contrast, when we stated in September 2007 that DNDO's Phase 1 tests were biased, we were referring to the test methods DNDO used, such as (1) using the same test sources and shielding materials during preliminary runs as were used during the actual tests, and (2) not using standard CBP operating procedures in testing the RIIDs.

We are recommending that the Secretary of DHS use the results of the Phase 3 tests solely for the purposes for which they were intended—to

identify areas needing improvement, not as a justification for certifying whether the ASPs represent a significant increase in operational effectiveness. However, if the Secretary of DHS intends to consider the results of the Phase 3 tests, along with other test data, in making a certification decision regarding ASPs, then we also recommend that the Secretary (1) direct the Director of DNDO to revise and clarify the Phase 3 test report to more fully disclose and articulate the limitations present in the Phase 3 tests—particularly the limitations associated with making comparisons between detection systems from a small number of test runs—and (2) clearly state which "relevant insights into important aspects of system performance" from the Phase 3 report are factored into any decision regarding the certification that ASPs demonstrate a significant increase in operational effectiveness. Finally, we further recommend that since there are several phases of additional ASP testing currently ongoing, the Secretary should direct the Director of DNDO take steps to ensure that any limitations associated with ongoing testing are properly disclosed when the results of the current testing are reported.

We provided DHS with a draft of this report for its review and comment. The department stated that it strongly disagreed with our draft report and two of our recommendations, it agreed to take some action on a third recommendation, and offered no comments on a fourth recommendation. In its comments, DHS cites narrative from the Phase 3 report explaining that the Phase 3 tests employed fewer test runs per test so as "to allow for more substantial variation among test cases" rather than "running sufficient number of repetitions… to provide high statistical significance results." Thus, in DHS's view, our assertion that the report does not "fully disclose" the Phase 3 test's limitations concerning the statistical significance of the results is incorrect. Our draft report recognizes DHS's description of how the Phase 3 tests were conducted. Our concern is that although DNDO cited the limited statistical significance of the test results at the outset of the Phase 3 report, it does not clearly state this limitation in expressing the test's findings. For example, as we note in our draft report, the Phase 3 report repeatedly states that the performances of the various ASPs were "statistically indistinguishable" even though DNDO did not perform enough test runs to estimate with a high degree of confidence whether the performances were actually similar. DNDO presents many of its findings as conclusive statements about ASP performance despite the fact that the Phase 3 test design cannot support these findings. DHS had additional comments, which are discussed at the end of this letter.

Background

In the summer of 2005, DNDO tested ASPs from 10 vendors to evaluate their performance capabilities and to select the ASPs that warranted further development and possible procurement. In July 2006, DNDO awarded contracts totaling $1.2 billion over fiveyears to three vendors—Raytheon, Canberra, and Thermo.

The Department of Homeland Security Appropriations Act for Fiscal Year 2007 states that "none of the funds appropriated … shall be obligated for full scale procurement of [ASP] monitors until the Secretary of Homeland Security has certified … that a significant increase in operational effectiveness will be achieved." Congress enacted a similar requirement in DHS's fiscal year 2008 appropriation. In hopes of obtaining secretarial certification by June 2007, DNDO tested ASPs at several sites, including the Nevada Test Site, the New York Container Terminal, the Pacific Northwest National Laboratory, and five ports of entry. DNDO conducted the tests at NTS in two phases. DNDO stated that the Phase 1 tests, performed in February-March 2007, attempted to estimate the performance capabilities of the ASPs with a high degree of statistical confidence. DNDO intended these tests to support the Secretary's decision on whether to certify the ASPs for the purposes of a full-scale production decision, while the Phase 3 tests were intended to help improve the computer algorithms that the ASPs use to identify the specific radiological or nuclear source inside a container.

On September 18, 2007, we testified that DNDO's Phase 1 tests did not constitute an objective and rigorous assessment of the ASPs' capabilities because, among other things, DNDO conducted preliminary test runs on source materials to be used in the tests, and then allowed the vendors to adjust their ASPs to specifically identify the source materials to be tested. We testified that in our view, DNDO's approach biased the tests in ways that enhanced the apparent performance of the ASPs. We also noted that the tests did not attempt to estimate the limits of ASPs' detection abilities—an important concern to those who will use them such as CBP officers. During that hearing, DNDO's Director stated that, contrary to statements DNDO made in its final Phase 3 test plan, DNDO would use the Phase 3 test results to help support the Secretary's decision on whether to certify the ASPs for full-scale production. Subsequently, DNDO delayed its anticipated date for secretarial certification to the fall of 2008 in order to conduct additional performance tests and field tests during fiscal year 2008.

DNDO's Phase 3 Test Report Frequently Overlooks the Limitations Associated with the Tests' Small Sample Sizes

Because the limitations of the Phase 3 test results are not properly discussed in the Phase 3 test report, the report does not accurately portray the results from the Phase 3 tests and could be misleading. The purpose of the Phase 3 tests was to identify areas in which the ASPs needed improvement. While some of the Phase 3 report addresses this purpose, much of the report compares the performance of the ASPs with each other or with PVTs and RIIDs during the tests. However, because DNDO performed each test a limited number of times, the data it used to make some of these comparisons provide little information about the actual capabilities of the ASPs. The narrative of the report often presents each test result as a single value, although, because of the limited number of test runs, the results would be more thoroughly and appropriately stated as a range of potential values. In addition, the report's narrative sometimes omits key facts that conflict with DNDO's analysis of the results.

The Phase 3 Test Report Largely Overlooks the Limiting Effects of Performing a Small Number of Tests

The purpose of the Phase 3 tests was to provide data that would help further develop and improve the ASPs' identification software and data to aid in the development of secondary screening operations and procedures. DNDO acknowledged early in the test report that the Phase 3 tests did not involve enough test runs to assess the performance of the ASPs with a high degree of statistical confidence:

> "The primary goals of the testing were to provide information by giving the ASP systems an opportunity to perform against a wider range of radionuclides, shielding, and cargo configurations. To allow for more substantial variation among test cases, the test plan called for a smaller number of trials over a larger range of objects and configurations rather than running sufficient number of repetitions for each individual test case to provide higher statistically significan[t] results." (p. 2)

DNDO also acknowledged early in the Phase 3 report that, given the small number of test runs, it would be difficult to compare the performance of the ASPs with each other:

> In these comparisons [of the performances of different ASP systems], results are often indistinguishable [i.e., not statistically significantly different] because the small sample sizes induce large uncertainties in the estimates of the probabilities being compared [for example: $n \leq 5$]."[4] (p.9)

[4]The report does not present significance-of-differences tests with its analyses.

Nonetheless, while some of the Phase 3 report addresses the stated purpose of the tests, the preponderance of the report compares the performance of the ASPs with each other or with PVTs or RIIDs during the tests, as shown in the following examples:

> "For [category of source material] at 2 mph, the ASP system performances are statistically indistinguishable." (p.13)

> "For shielded [category of source material], performance for all three systems is statistically indistinguishable with probabilities of correct alarm varying approximately between 0.84 and .0.92." (p.11)

The statements imply that the performances of the ASPs were similar because the results were "statistically indistinguishable." However, given the small number of test runs, it is impossible to determine with a high degree of confidence whether or not the performances were actually similar. Yet the report's text describing specific results rarely qualifies the results by stating that the test was run only a few times or that the results should not be considered conclusive of the ASPs' capabilities.

Similarly, the report's executive summary presents the test results as conclusions about the performance capabilities of the ASPs, PVTs, and RIIDs:

> "For the source configurations tested, the ASP systems have equal performance down to the lowest source activity tested." (p. iii)

> "The PVT systems display lower performance than the ASP systems for [category] and [category] sources." (p. iv)

> "When comparing the ASP systems _ mph identification metric with the _ RIID measurements…, it is observed that the RIID performance is poor compared to the ASP systems." (p. iv)[5]

[5]DNDO's Phase 3 report is classified. Because of this, some of the quotes from the report are missing information in order to protect sensitive information.

Report Text Often Omits the Range of Values Surrounding Each Test Result

The Phase 3 test report makes some of its performance comparisons by citing the percentage of correct detections or identifications that an ASP made on a test. For example:

> "For bare [category of source material] only, … [T]he probability of correct identification [for the 3 ASPs] varied between 0.34 and 0.5." (p.14)[6]

However, because each test involved a small number of test runs, these percentages provide little information about the performance capabilities of the ASPs. In fact, because of the small number of test runs, DNDO can only estimate that each ASP can correctly identify the type of source material within a range of values.[7] The fewer the number of test runs, the larger the range. For example, for the ASP that correctly identified the source material 34 percent of the time during the tests, the report text omits the fact that, as shown on an accompanying graph, DNDO can only estimate that the ASP would be able to correctly identify the source between about 10 percent and 65 percent of the time. By stating that the ASP identified the source 34 percent of the time without clarifying that the results came from only a few test runs, the report's text makes the test results seem more precise than they really are. Similarly, for the ASP that correctly identified the source material 50 percent of the time during the tests, the small number of test runs means that DNDO can only estimate that the ASP would be able to correctly identify the source material between about 15 percent and 85 percent of the time. This range is too wide to have much value in determining how well the ASP may perform in the real world. Although these ranges are clearly shown on the report's graphs, they are omitted in the report's descriptions and interpretations of the test results.

Similarly, DNDO's analysis comparing the performances of ASPs and RIIDs fails to consider the uncertainties created by the tests' small sample sizes. The report states that the RIIDs "performance is poor compared to the ASP systems." For example, during the tests, one vendor's ASP correctly identified one type of source material about 50 percent of the time, while the RIIDs correctly identified the same type of source material

[6]DNDO analyzed this series of tests by source category (medical, industrial, or special nuclear material) rather than by specific source material or isotope. For ease of discussion, we refer to each category as a source material.

[7]Unless stated otherwise, the range of values represents the confidence intervals surrounding the point estimate at the 95 percent level.

about 10 percent of the time. However, given the small number of test runs, DNDO cannot be confident that these percentages precisely indicate the performance capabilities of the ASPs and RIIDs. On the basis of the tests, DNDO can only infer that the ASPs' and RIIDs' performance capabilities lie somewhere within a relatively large range of values. As these ranges are illustrated in the report's graphs, it appears that the difference in the performance of the ASPs and RIIDs may not be statistically different for three of the five types of source materials DNDO tested. This does not necessarily mean that the ASPs and RIIDs performed equally well; rather, DNDO did not conduct each test enough times to determine that the superior performance of the ASPs over the RIIDs reflects the capabilities of the ASPs rather than mere chance.

DNDO's Phase 3 Test Report Omits Important Details that Affect the Interpretation of the Test Results

In a few instances, the report omits important details concerning DNDO's interpretation of the results. For example, DNDO seems to assert in the report's executive summary that the ASPs are as good as the PVTs at detecting radiological or nuclear source materials:

> "The ASP systems demonstrated detection limits equal to or better than those of any of the PVT systems as configured during testing." (p.iii)

However, the report's executive summary fails to note that because DNDO used only one type of source material, the results are largely specific to that particular source material and would not necessarily apply to other specific source materials. In fact, for other types of source material, the report shows several instances in which the PVTs were apparently able to detect other types of source materials better than the ASPs. Moreover, other Phase 3 tests showed that simply moving the source material referred to in the above quote to another place in the container affected the relative performances of the ASPs and PVTs.

Similarly, in reporting how well the ASPs performed when the radiation from the source material was partially blocked by a shielding material, DNDO stated:

> "the ASP systems have the ability to identify sources when placed inside almost all but the thickest shielding configuration tested." (p.iv)

Again, however, DNDO fails to note in its report that, as it explained in its Phase 3 test plan, all the shielding used in the Phase 3 tests represented "light shielding." The report also fails to state how many specific sources

the ASPs could correctly identify or how frequently the ASPs could identify them.

Phase 3 Test Results Provide Little Evidence as To Whether ASPs Represent an Improvement Over Currently Deployed Technology

In our view, it is not appropriate to use the Phase 3 test report in determining whether the ASPs represent a significant improvement over currently deployed radiation equipment because the limited number of test runs does not support many of the comparisons of ASP performance made in the Phase 3 report. As noted, DNDO's use of a small number of runs for each test means that DNDO can only be certain that the ASP can correctly identify or detect a source material over a broad range of possible values rather than at a specific rate. This is true even if the ASP was successful every time a test was conducted. For example, as noted in the Phase 3 test report, if the ASP correctly identified a source material 100 percent of the time, but the test was run only five times, the most DNDO can estimate is that the ASP should be able to correctly identify the source no worse than about 60 percent of the time.

The Phase 3 test results do not help to determine an ASP's "true" level of performance because DNDO did not design the tests to assess ASP performance with a high degree of statistical confidence. In the Phase 3 test plan, DNDO was very clear that it had intended the tests to help develop a conduct of operations for secondary screenings and to cover a larger array of source materials and test scenarios than were conducted in the Phase 1 tests.

DNDO also originally stated that the Phase 3 tests would not be used for secretarial certification that the ASPs represented a "significant operational improvement" over currently deployed radiation detection equipment. DNDO stated that it had designed the Phase 1 tests to "evaluate the current state of performance of the ASP…systems." However, prior to releasing the Phase 3 report, DNDO changed its position, stating in the final Phase 3 test report that the test results are relevant to secretarial certification:

> "The Phase 3 test campaign was not originally intended to support the Secretarial Certification of the ASP systems. However, the test results provide relevant insights into important aspects of system performance and should be taken into consideration by the Secretary of Homeland Security in making his (ASP procurement) decision." (p.iii)

It is important to note that DNDO does not elaborate in the test report as to what the "relevant insights" are or how they relate to Secretarial

certification. DNDO also does not explain why those insights would be relevant considering that, as stated in the Phase 3 test plan, the results from the tests lack a high degree of statistical significance. Finally, it should be noted that when the Director of DNDO testified in September 2007 that the Phase 3 test results would help inform the Secretary's recommendation, he also acknowledged that the Phase 3 test report had not yet been prepared.

SNL's Special Tests Were Designed to Improve ASP Performance

The special tests were performed by experts from Sandia National Laboratories who were not part of the Phase 1 or Phase 3 tests. The special tests were designed to examine potential vulnerabilities associated with either the ASPs or the Phase 1 or Phase 3 test plan and vulnerabilities in DNDO's test processes. Conducting this type of test would allow the ASP vendors the opportunity to make improvements to their systems in order to address weaknesses revealed during the special tests. Like the Phase 3 tests, the special tests used a small number of runs for each testing scenario. Because of the small number of runs, the test results do not support estimating the probability of detection or identification with a high confidence level making it difficult to use the results of the special tests to support a certification decision by the Secretary of DHS. On this point, the special test report acknowledges that "the special tests were never intended to demonstrate conformity of the [ASP] systems against specific performance requirements."

From the special tests, SNL concluded

1. "Areas for software and hardware improvement have been identified based on system performance issues observed for the versions of the ASP hardware and software tested at the NTS during Winter 2007."

2. "For the data made available to us, the reported results … are consistent with the underlying collected data—indicating that the DNDO ASP system assessment was not biased."

3. "Recommendations to improve the testing rigor have been made…(noting that) their implementation must be balanced against other test campaign impacts (such as) cost, schedule, availability of resources, etc.," and

4. "Based on our limited tests we observed no data suggesting that the ASP system performance was inappropriately manipulated by either the vendors or the test team."

Overall, the special test report appears to accurately describe the purpose, limitations, and results of the special tests. In our view, DNDO should consider SNL's views as it proceeds with additional ASP testing in 2008. It is important to note, however, in Sandia's conclusions that the "ASP system assessment [in 2007] was not biased" and that it "observed no data suggesting that the ASP system performance was inappropriately manipulated," Sandia is referring to the data derived from ASP tests. However, SNL does not comment on the biased testing methods we identified during the Phase 1 ASP tests at the Nevada Test Site in 2007. Specifically, when we stated in September 2007 that DNDO's Phase 1 tests were biased, we were referring to DNDO's test methods which (1) used the same test sources and shielding materials during preliminary runs as were used during the actual tests and (2) did not use standard CBP operating procedures in testing the RIIDs.

Conclusions

Preventing the material for a nuclear weapon or a radiological dispersal device from being smuggled into the United States remains a key national security priority. Testing radiation detection equipment to understand its capabilities and limitations is an important part of preventing nuclear smuggling. The Phase 3 and special tests were part of DNDO's 2007 effort to test ASPs in order to identify areas for further development to these devices. The Phase 3 test results are relevant to DNDO's original objective for the Phase 3 tests—to identify areas in which the ASPs needed improvement. However, because of the limitations of the tests, DNDO should not be using the test results as indicators of the overall performance capabilities of the ASPs. Moreover, in the Phase 3 report, DNDO presented and analyzed the test results without fully disclosing key limitations of the tests, which is not consistent with basic principles of statistics and data analysis. Because of this, many of the report's presentations and comparisons of performance among ASPs and between ASPs and PVTs are not well supported and are potentially misleading. Regarding the special tests, SNL notes in its test report that it designed the tests to identify areas where the ASPs need to improve—not to measure the ASPs performance against requirements. Overall, because of the limitations discussed in this report, it is our view that neither the Phase 3 tests nor the special tests should serve as a basis for the Secretary of DHS whether the ASPs represent "a significant increase in operational effectiveness" over current radiation detection equipment.

Recommendations for Executive Action

To ensure that the limitations of the Phase 3 test results, and future ASP test results, are clearly understood, we are making the following four recommendations.

We recommend that the Secretary of DHS use the results of the Phase 3 tests solely for the purposes for which they were intended—to identify areas needing improvement, not as a justification for certifying whether the ASPs warrant full-scale production.

However, if the Secretary of DHS intends to consider the results of the Phase 3 tests, along with other test data information, in making a certification decision regarding ASPs, then we recommend that the Secretary take the following actions:

- Direct the Director of DNDO to revise and clarify the Phase 3 test report to more fully disclose and articulate the limitations present in the Phase 3 tests—particularly the limitations associated with making comparisons between detection systems from a small number of test runs.
- Clearly state which "relevant insights into important aspects of system performance" from the Phase 3 report are factored into any decision regarding the certification that ASPs demonstrate a significant increase in operational effectiveness.

Finally, we further recommend that since there are several phases of additional ASP testing currently ongoing, the Secretary should direct the Director of DNDO take steps to ensure that any limitations associated with ongoing testing are properly disclosed when the results of the current testing are reported.

Comments from the Department of Homeland Security and Our Evaluation

We provided DHS with a draft of this report for its review and comment. Its written comments are presented in appendix I. The department stated that it strongly disagreed with our draft report and two of our report's recommendations. DHS agreed to take some action on a third recommendation and offered no comments on a fourth recommendation. The department stated several reasons for its disagreement. First, DHS cites narrative from the Phase 3 report explaining that the Phase 3 tests employed fewer test runs per test so as "to allow for more substantial variation among test cases" rather than "running sufficient number of repetitions … to provide high statistical significance results." Thus, in DHS's view, our assertion that the report does not "fully disclose" the Phase 3 tests' limitations concerning the statistical significance of the

results is incorrect. Our draft report recognizes DHS's description of how the Phase 3 tests were conducted. Our concern is that although DNDO cited the limited statistical significance of the test results at the outset of the Phase 3 report, DNDO's findings do not reflect this limitation. For example, as we note in our draft report, the Phase 3 report repeatedly states that the performances of the various ASPs were "statistically indistinguishable" even though DNDO did not perform enough test runs to estimate with a high degree of confidence whether the performances were actually similar. DNDO presents many of its findings as conclusive statements about ASP performance despite the fact that the Phase 3 test design cannot support these findings.

Second, the department commented that the Phase 3 test report clearly and succinctly stated another limitation of the test methodology—specifically, that the tests were not designed to be a precise indicator of ASP performance. In the department's view, noting this limitation throughout the Phase 3 report would have been unwieldy. We did not expect DNDO to repeat this limitation throughout the report. However, as suggested in our report, the Phase 3 report should accurately reflect the test results without portraying the results as being more precise than they really are. Using an example from the Phase 3 report, if DNDO notes that an ASP successfully identified a specific source material 34 percent of the time during the tests, it should also indicate that, given the small number of test runs, DNDO can only estimate that the ASP would be able to correctly identify the specific source material between 10 and 65 percent of the time. However, no such discussion of the wide range of potential results is included in the report's narrative. In our view, presenting the test results without sufficient narrative about the tests' limitations is potentially misleading.

Third, the department stated that although the Phase 3 tests were not intended to support the DHS Secretary's certification decision, DHS decided that it needed to consider all available test results in making this decision. DHS further commented that not doing so would subject it to criticism of "cherry-picking" the results. In response, although we acknowledge the need to consider all available test results, we believe they should be considered in their appropriate context, and that test results do not all carry the same weight. In our view, test results with a high degree of statistical significance (i.e., unlikely to be the result of chance) should be considered a better indicator of ASP performance than those with a lower level of statistical significance. Because the Phase 3 tests involved only 1-10 runs per test, very few of the results can be generalized as reliable estimates of how the ASPs perform and thus

potentially provide questionable evidence for the certification process. We also note that, in its comments, DHS did not address what Phase 3 results or important insights it considered to be relevant to Secretarial certification.

Fourth, DHS comments that our draft report failed "to acknowledge the depth and breadth of the ASP test campaign, which is by far the most comprehensive test campaign ever conducted on radiation detection equipment." However, our report describes previous ASP testing and some of our prior findings about that testing, and notes that ASP testing continues in 2008. More importantly, the extent of testing is not the issue at hand. In our view, regardless of how many tests are performed, the tests must employ sound, unbiased methodologies and DNDO should draw and present conclusions from the test results in ways that accurately and fully reflect the data and disclose their limitations.

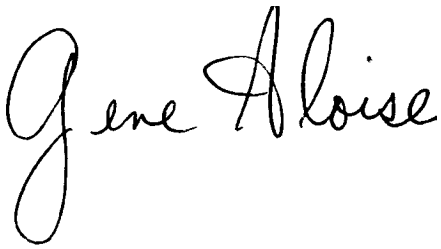
DHS stated that it disagreed with our recommendations to (1) use the Phase 3 test's to identify areas needing improvement and not as a basis for certification and (2) revise and clarify the Phase 3 report to reflect the limitations in the tests' methodology and results. It did not offer comments on our recommendation that the Secretary clearly state what relevant insights from the Phase 3 report are factored into any certification decision. We continue to believe that the Phase 3 tests should be used only for the intended purpose stated in its test plan—to improve the software of ASPs. We would also note that our draft report recommends that DNDO revise and clarify the Phase 3 test report only if it includes Phase 3 test results among the data that will be presented to the Secretary prior to his decision on certification. If DNDO chooses to use the Phase 3 test results for certification, we believe it is important that DNDO explain what test results are relevant to certification and why the value of those results are not mitigated by the limitations associated with the Phase 3 tests' small sample sizes.

In response to our last recommendation, the department stated that it has taken and will continue to take steps to ensure that it properly discloses any limitations associated with ongoing testing as it moves toward secretarial certification of the ASPs.

As agreed with your offices, unless you publicly announce the contents of this report, we plan no further distribution until 30 days from the report date. At that time, we will send copies of this report to the Secretary of DHS and interested congressional committees. We will also make copies

available to others upon request. In addition, this report will be available at no charge on the GAO Web site at http://www.gao.gov.

If you or your staffs have any questions about this report, please contact me at (202) 512-3841 or aloisee@gao.gov. Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. GAO staff that made major contributions to this report are listed in appendix II.



Gene Aloise
Director,
Natural Resources and
Environment

Appendix I: Comments from the Department of Homeland Security

U.S. Department of Homeland Security
Washington, DC 20528

Homeland Security

August 29, 2008

Mr. Gene Aloise
Director, Natural Resources and Environment
Government Accountability Office
441 G Street NW
Washington, DC 20548

Dear Mr. Aloise:

Re: Draft Report GAO-08-979, Combating Nuclear Smuggling: DHS's Phase 3 Test Report on Advanced Portal Monitors Does Not Fully Disclose the Limitations of the Test Results.

The Department of Homeland Security strongly disagrees with the GAO Draft Report. The title of the GAO report "DHS's Phase 3 Test Report on Advanced Portal Monitors Does Not Fully Disclose the Limitations of the Test Results" is misleading and not substantiated by the body of the GAO report. First, the Test Purpose section of the test report clearly states that "to allow for more substantial variation among test cases, the test plan called for a smaller number of trials over a larger range of objects and configurations rather than running sufficient number of repetitions for each individual test case to provide high statistical significance results."[1] This statement applies to the entire test report; therefore to say that DHS did not "fully disclose" limitations that could be "potentially misleading" is false. Repeating the same statement throughout the test report would be unnecessary and tedious.

Second, DHS clearly acknowledges that the main purpose of the Phase 3 test was to provide information to improve the system algorithms.[2] The ASP Phase 3 test was designed to identify areas for ASP development and improvement. The test was not intended to be a precise indicator of ASP performance, nor does the test report ever claim to draw such conclusions. The underlying test design purposely involved test cases that were different from those against which ASP performance is to be measured. The test report clearly and succinctly stated the limitations associated with the test methodology and analysis approach. Again, given the amount of

[1] "Test Report in Support of Advanced Spectroscopic Portal (ASP) Systems Development at the Nevada Test Site (NTS)", Section 1.2 Test Purpose, Page 2
[2] "Test Report in Support of Advanced Spectroscopic Portal (ASP) Systems Development at the Nevada Test Site (NTS)", Section 1.2 Test Purpose, Page 2 and Section 2 Test Results, Page 4

www.dhs.gov

material and density of the test report, it would have been unwieldy to repeat caveats throughout the report.

Third, even though the Phase 3 report was not designed to support the certification decision or to substantiate performance of the systems against the defined threat basis, DHS quickly recognized that it needed to consider all available test results in making a certification decision. For that reason, the Phase 3 Test report clearly states "Although the Phase 3 test campaign was not originally intended to support the secretarial certification of the ASP systems, the test results provide relevant insights into important aspects of the system performance and should be taken into consideration by the Secretary of Homeland Security."[3] Indeed, had we not done so, DHS would now be accused of "cherry-picking" test results, or worse, ignoring data.

Once again, by reporting on ASP testing in a piece-meal fashion, GAO fails to acknowledge the depth and breadth of the ASP test campaign, which is by far the most comprehensive test campaign ever conducted on radiation detection equipment. The ASP Phase 3 test is but one in a series of carefully-designed tests conducted over a period of years in the path to secretarial certification and full-rate production. To date, these campaigns have included a developmental set of testing for the engineering developmental models in Winter 2007 and performance testing at the Nevada Test Site (NTS), deployment readiness testing at the Pacific Northwest National Laboratory (PNNL), operational testing at the New York Container Terminal (NYCT), and field validations at multiple POEs conducted in Summer 2008.

In conjunction with the Department of Energy (DOE), DNDO is also conducting Threat Injection Studies for ASP systems. These threat injection studies will examine the limits of performance of ASP in order to guide the setting of thresholds and concept of operations (CONOPS) and also compare ASP and polyvinyl toluene (PVT) performance. To date, DNDO has developed, integrated, and validated a set of tools to perform the injection studies using a standard data format. Data collection has also been underway to create threat-representative signatures that can be injected into stream-of-commerce data. In preparation for the studies, DNDO has worked to collect data to validate injection methodology and prepare data set of approximately 8000 validated stream-of-commerce data files into which threat signatures will be injected and also create threat-representative signatures based on the collected data.

Additional 2008 ASP testing is currently underway, and includes: system qualification testing (SQT) to verify compliance with the ASP performance specification; integration testing at PNNL to verify that ASP performance remains sound when it is integrated into the POE architecture; performance testing at NTS to validate the detection and identification capabilities of ASP systems against special nuclear materials (SNM) and materials for radiological dispersal devices (RDD); and operational test and evaluation activities to validate operational performance of the system at POEs. The successful completion of these steps will provide data for the Secretary's Certification decision. DNDO will use a combination of cost-benefit analyses as well as demonstrated performance metrics to assist in the Secretary's certification decision. Part of the certification process will involve working with the National Academy of Sciences, to review DNDO test plans and procedures, as required in the FY 2008 Consolidated Appropriations Act.

[3] "Test Report in Support of Advanced Spectroscopic Portal (ASP) Systems Development at the Nevada Test Site (NTS)", Section 1.2 Test Purpose, Page 2 and Executive Summary, Page iii

DHS will also use the test results, along with other information, to seek approval of the DHS Investment Review Board (IRB) prior to proceeding to full-scale production and deployment of ASP systems at POEs.

Based on the above, DHS believes that the Phase 3 Test Report more than adequately discloses the purpose and limitations associated with the Phase 3 test results, and therefore, disagrees with the GAO's recommendation that the Phase 3 report needs to be revised and clarified. DHS also believes that all data needs to be considered for the secretarial certification decision, and therefore, disagrees with the GAO's recommendation that the Phase 3 data not be considered. It is neither prudent nor scientifically justifiable to pick the data one chooses to use for making a decision. While some data may carry more weight than others (particularly the most recent data for the newest software version), no data should be ignored. DHS has taken, and will continue to take, steps to ensure that any limitations associated with ongoing testing are properly disclosed in the path to certification.

Thank you for the opportunity to review and provide comments to the draft report.

Sincerely,

Jerald E. Levine

Jerald E. Levine
Director, Departmental GAO/OIG Liaison Office

Appendix II: GAO Contact and Staff Acknowledgments

GAO Contact

Gene Aloise, (202) 512-8051 or aloisee@gao.gov

Staff Acknowledgements

In addition to the contact named above, Ned Woodward, Assistant Director; James Ashley, Nabajyoti Barkakati, Carol Kolarik, Omari Norman, Alison O'Neill, Anna Maria Ortiz, Daren Sweeney, Michelle Treistman, and Gene Wisnoski made significant contributions to this report.