This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.
New Tools for Assessing State and Local Capabilities for Countermeasure Delivery

Christopher Nelson, Edward W. Chan, Carol Fan, Debra Lotstein, Leah B. Caldarone, Shoshana R. Shelton, Amy L. Maletic, Andrew M. Parker, Alexandria Felton, Amanda Pomeroy, Elizabeth M. Sloss

Prepared for the Department of Health and Human Services
This work was sponsored by the U.S. Department of Health and Human Services Office of the Assistant Secretary for Preparedness and Response. The research was conducted in RAND Health, a division of the RAND Corporation.

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world. RAND’s publications do not necessarily reflect the opinions of its research clients and sponsors.

RAND® is a registered trademark.

© Copyright 2009 RAND Corporation

Permission is given to duplicate this document for personal use only, as long as it is unaltered and complete. Copies may not be duplicated for commercial purposes. Unauthorized posting of RAND documents to a non-RAND Web site is prohibited. RAND documents are protected under copyright law. For information on reprint and linking permissions, please visit the RAND permissions page (http://www.rand.org/publications/permissions.html).
Eight years after the anthrax attacks of October 2001, the United States lacks a set of clear and defensible measurement tools for assessing the nation’s public health emergency preparedness. Countermeasure delivery -- the ability to dispense life-saving medications and other medical supplies to large populations, often on very short timelines -- is one area in which considerable progress has been made. Even in this area, measurement tools have focused on assessing the existence of personnel, plans, and equipment, not on whether the health system can put these resources into practice in emergency conditions.

The United States Department of Health and Human Services (HHS) asked RAND to assist the Centers for Disease Control and Prevention, Division of the Strategic National Stockpile (CDC/DSNS) in developing tools for assessing the readiness of state and local health departments to carry out countermeasure-delivery operations. This report describes those tools and provides an approach for measure development for public health emergency preparedness that can be applied more broadly. The operations manuals that are discussed in this report provide guidance on assessing a jurisdiction’s countermeasure-delivery capabilities. The assessments are based on a series of small-scale drills, each focusing on a specific building-block capability and on specific metrics to be collected in each drill. These tools were field-tested in several state and local health departments. The manuals and the assessments are available on the CDC website at: http://emergency.cdc.gov/cotper/coopagreement/.

This report will be of interest to state and local health departments that are responsible for planning and exercising countermeasure delivery, as well as to other agencies charged with supporting countermeasure-delivery operations. It will also be of interest to those who want a deeper description of the thought process that went into the assessments included in the CDC cooperative-agreement guidance. In addition, the report will be of use to those interested in broader exercise and measurement issues.

Throughout this report and the operations manuals, we use the term “patient” to refer to the individuals that visit PODs to receive countermeasures. We chose this term to be consistent with the language used in the POD modeling programs referenced in the Dispensing Assessment Manual. In an actual emergency, PODs would be expected to serve healthy, non-injured persons, who would not be considered patients in the strict medical sense. True patients, i.e., sick or injured individuals, would go to healthcare facilities for treatment instead.

This work was sponsored by the U.S. Department of Health and Human Services Office of the Assistant Secretary for Preparedness and Response (HHS/ASPR) and was executed by the Center for Public Health Preparedness within RAND Health. A profile of the Center for Public Health Preparedness, abstracts of its publications, and ordering information can be found online (http://health.rand.org/preparedness).
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREFACE</td>
<td>i</td>
</tr>
<tr>
<td>FIGURES</td>
<td>iii</td>
</tr>
<tr>
<td>TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>v</td>
</tr>
<tr>
<td>ABBREVIATIONS</td>
<td>x</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>xi</td>
</tr>
<tr>
<td>Chapter 1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 2. Determining What to Assess</td>
<td>4</td>
</tr>
<tr>
<td>Chapter 3. Overview of Operations Manuals</td>
<td>12</td>
</tr>
<tr>
<td>Chapter 4. Field Testing: Methods and Results</td>
<td>21</td>
</tr>
<tr>
<td>Chapter 5. Implications and Next Steps</td>
<td>31</td>
</tr>
<tr>
<td>APPENDIX A: Analysis Used to Select SNS Elements Around Which to Build Assessments</td>
<td>36</td>
</tr>
<tr>
<td>APPENDIX B: Questions for Eliciting Feedback on Field Tests of SNS Capability Assessments</td>
<td>39</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>41</td>
</tr>
</tbody>
</table>
FIGURES

Figure 2.1: Conceptual Model of Public Health Emergency Preparedness and Response
Figure 2.2: Capabilities Selected for Assessment Tool Development
Figure 2.3: Crosscutting and Function-Specific Capabilities can be Combined into Compound Drills
Figure A.1: Precedence Diagram with Crosscutting Capabilities Highlighted
Figure A.2: Precedence Diagram with Function-Specific Capabilities Highlighted
TABLES

Table 3.1: Content of Operations Manuals........................................................................ 12
Table 3.2: Overview of the Current Set of Operational Assessments................................. 14
Table 4.1: Field Tests Conducted.................................................................................... 23
SUMMARY

If a biological attack occurs or an infectious disease breaks out, medical countermeasures, such as vaccines, antibiotics, or antidotes, may need to be delivered rapidly to the general public. The federal government, through the Centers for Disease Control and Prevention (CDC), maintains a cache of pharmaceutical drugs and other medical materials, known as the Strategic National Stockpile (SNS). Those materials [called materiel in the main text] can be sent to state and local public health departments. State and local governments are responsible for requesting the material, receiving it from the federal government, distributing it to local areas, and, ultimately, dispensing countermeasures to members of the public.

Assessing state and local departments’ readiness to perform these functions is a critical element for improving their ability to conduct a countermeasure-delivery campaign in an emergency. The CDC’s Division of the Strategic National Stockpile (CDC/DSNS) has been using a checklist-based approach, known as the Technical Assistance Review (TAR), to assess the jurisdictions’ countermeasure delivery plans, equipment, and personnel. However, a consensus has emerged within the U.S. Department of Health and Human Services (HHS) and CDC that assessing these capacity elements is not sufficient; methods are needed for assessing the jurisdictions’ capabilities to implement and adapt these plans in real-world situations.

To address this need, the United States Department of Health and Human Services Office of the Assistant Secretary for Preparedness and Response (HHS/ASPR) asked RAND to assist CDC/DSNS in developing a system for assessing the operational capabilities to engage in mass countermeasure delivery.

This report describes the assessments that were developed, and it provides an approach for measuring development for public health emergency preparedness that can be applied more broadly. The operations manuals that are discussed in this report provide guidance on assessing a jurisdiction’s countermeasure-delivery capabilities. Users can download and print spreadsheet-based data-reporting tools that will facilitate calculation of required metrics and collection of key data. The manuals and the data-reporting tools are available on the CDC website at: http://emergency.cdc.gov/cotper/coopagreement/.

Throughout this report and the operations manuals, we use the term “patient” to refer to the individuals that visit PODs to receive countermeasures. We chose this term to be consistent with the language used in the POD modeling programs referenced in the Dispensing assessment manual. In an actual emergency, PODs would be expected to serve healthy, non-injured persons, who would not be considered patients in the strict medical sense. True patients, i.e., sick or injured individuals, would go to healthcare facilities for treatment instead.
ASSESSMENT TOOLS FOCUS ON “BUILDING-BLOCK” CAPABILITIES

Given the rarity of large-scale public health emergencies, there is little evidence base to guide decisions about what aspects of preparedness are most important to measure. Thus, we conducted engineering-style process analysis to identify core “building-block” capabilities that can be deployed, combined, and adapted in response to a broad spectrum of response scenarios. Focusing assessments around a small number of critical operational capabilities helps keep assessment burdens reasonable and promotes flexibility by encouraging jurisdictions to think of situation-specific plans as containing certain common elements. Assessment tools were ultimately created for five building-block capabilities:

- **Staff call-down**: the capability to notify and mobilize staff to perform emergency response functions.
- **Site call-down**: the operational capability to notify facilities (e.g., points of dispensing [PODs] or warehouses) that are not ordinarily under the health department’s control and determine how quickly the sites could be made ready for emergency operations.
- **Warehouse pick-list generation**: a jurisdiction’s ability to use its warehouse inventory system to ensure a timely match between the demand for and supply of critical resources.
- **POD set up**: the ability to quickly set up a facility for use as a POD.
- **Mass dispensing through PODs**: the ability to quickly deliver medications or other “countermeasures” to large numbers of people at a POD.

The assessments for each of these capabilities are structured in the form of an operational exercise. Each assessment features a set of clearly defined performance metrics. The Homeland Security Exercise Evaluation Program (HSEEP) and other documents provide a considerable amount of guidance on exercise evaluation; however as yet there are no comparable performance metrics that can support reliable cross-sectional and temporal comparisons of performance of key capabilities during exercises. This work is intended to meet that need.

To reduce assessment burdens, we designed each of the assessments to be used in a small-scale drill that can be conducted individually or combined as components of larger exercises. This versatility increases the range of exercises on which the assessments can be used and decreases the need for additional exercises. The assessments can also be embedded in routine public health functions, thus reducing the need to assemble key staff solely for exercising.

For each assessment, an operations manual was created with a summary of the assessed capability and detailed instructions on drill design, data collection, and calculating and reporting metrics.
FIELD TESTS ASSESSED THE FEASIBILITY, UTILITY, AND RELIABILITY OF ASSESSMENTS

Four of the five assessment tools were field-tested in 2007–2008 (one of the assessments, site call-down, was not field tested separately, because its structure and mechanics are nearly identical to the assessment of staff call-down, which was tested). In nearly all cases, field tests were done in conjunction with already-scheduled drills being conducted by the state or local health department. Following the exercise, RAND staff used a semi-structured interview protocol to guide discussions with members of the health department about the feasibility, reliability, validity, and utility of the assessment tools. Key findings were as follows:

- **Assessments were useful and feasible.** In general, field-test sites reported that the metrics provided useful feedback, that the manuals and spreadsheet-based tools were useful for managing data, and that the level of effort and skill required to use them was reasonable. However, the assessment of the pick-list generation required a higher level of assistance and explanation, largely because the field-test sites were unfamiliar with the concept of a drill devoted solely to exercising the pick-list-generation function.

- **Time studies and computer models presented special challenges.** The mass dispensing assessment requires sites to conduct time studies of POD stations to measure the number of minutes or seconds required to process patients. While collecting the information on timing was feasible for sites during the field tests, staff expressed concern about the number of people needed to do the timing. Moreover, staff at the field-test site were generally not comfortable with computer modeling.

- **Flexibility built into assessments might limit data comparability across sites.** The assessments provide considerable flexibility to sites in specifying exercise scenarios and response processes. However, such variations might significantly limit the ability to draw valid comparisons among performance-metric data across sites and over time. For instance, allowing health departments to choose their own scenario parameters (e.g., time of day for staff call-down, patient mix in the mass-dispensing assessment, start and stop times for POD set up) affected observed performance and made it nearly impossible for us to compare performance on the same drill across pilot sites.

- **More testing is required to assess validity.** Those in the test sites generally found that the results of the assessments had face validity. However, we were unable to validate the results from these assessments against results from real-world responses or other assessments of the same capability.

IMPLICATIONS AND NEXT STEPS

Policymakers must consider several issues when implementing the assessments.
Further Refinements Will Require Decisions About Balancing Standardization and Flexibility

Findings from the field tests suggest that further efforts will be required to control sources of variation captured by the assessments that go beyond the operational capabilities they were designed to measure (also known as error variance\(^1\)). Two possible approaches to address these issues are (a) to prescribe more specifically the drill scenarios and tasks, and (b) to collect information on drill context and processes and use this information to make post-hoc adjustments in interpreting assessment results.

Given that a key goal in designing the assessments was to maximize flexibility to jurisdictions, the assessments generally favored the second option (b), collecting extensive information on drill scenarios and procedures. However, the success of any post-hoc adjustment strategy will depend on the number of jurisdictions reporting data and the ability to identify the right contextual variables. As a result, post hoc adjustment might need to be supplemented by greater standardization in scenarios and procedures.

Consider Technical-Assistance Requirements of Performance Assessments

Most of the assessments appear to be quite feasible; however, it is likely that training will be required to ensure consistent application of the procedures and metrics provided in the manuals. Given that expertise with computer modeling varies among health departments, the computer-based dispensing assessment will likely require substantially more technical assistance than the others.

Recommended Next Steps

While making plans to roll out the assessment tools described in this report, HHS/ASPR and CDC/DSNS should consider the following recommendations:

The results of the field testing highlight that the science and art of measuring public health emergency response capabilities remain in their infancy. The assessments were found to be useful and feasible; yet, it is clear that data collected across different sites might not be fully comparable. Even imperfect assessments of capability provide an important complement to the current practice of measuring preparedness by plans, staffing, and equipment. Thus, it seems reasonable to consider rolling out the assessments on a trial basis.

However, decision makers should ensure that processes are in place to capture lessons learned from initial rollout and revise the measures accordingly. Specifically, they must

- **Collect performance-metrics data during the first year of implementation.** Such data will assist in the process of refining the measures. CDC/DSNS may want to consider developing a standardized feedback form to capture lessons from awardees.

---

\(^1\) Variance arising from unrecognized or uncontrolled factors in a study.
• **Analyze data collected during the first year and revise assessments accordingly.** Analyzing data from initial implementation will help assess the quality of the data. Data distributions should be examined for unexpected variation, with the goal of distinguishing true score variance from that due to errors in conducting the assessment. Assessment manuals and metrics should be revised to address problems revealed by this analysis.

• **Develop strategies for ensuring that awardees use the metrics to improve performance.** CDC should clarify how the data from the assessments might fit into the Pandemic and All-Hazards Preparedness Act’s (PAHPA’s) penalties for nonperformance and pay-for-performance requirements and consider a more focused pilot study to develop best practices in process improvement.

**Continue to develop the evidence base behind measures.** The data collected for these metrics will provide a valuable tool for more in-depth studies. For example, the data might be used to examine whether capacity measures on the Technical Assistance Review predict actual performance on operational-capability metrics.
ABBREVIATIONS

AAR       after-action report
AHRQ      Agency for Healthcare Research and Quality
ASTHO     Association of State and Territorial Health Officials
ASPR      Office of the Assistant Secretary for Preparedness and Response, US
          Department of Health and Human Services
BERM      Bioterrorism and Epidemic Outbreak Response Model
CBP       capabilities-based planning
CDC       Centers for Disease Control and Prevention, an agency of HHS
CRI       Cities Readiness Initiative
DSLR      Division of State and Local Readiness, a division within the Centers for
          Disease Control and Prevention, Coordinating Office for Terrorism
          Preparedness and Emergency Response
DSNS      Division of the Strategic National Stockpile, a division within Centers
          for Disease Control and Prevention, Coordinating Office for Terrorism
          Preparedness and Emergency Response
EEG       Exercise Evacuation Guide
EOC       Emergency Operations Center
FY        fiscal year
FEMA      Federal Emergency Management Agency
HHS       US Department of Health and Human Services
HSEEP     Homeland Security Exercise Evaluation Program
HSPD-8    Homeland Security Presidential Directive 8
IOM       Institute of Medicine
MOU       Memorandum of Understanding
NACCHO    National Association of City and County Health Officials
NIMS      National Incident Management System
PAHPA     Pandemic and All-Hazards Preparedness Act
PHEP      Public Health Emergency Preparedness
POD       point of dispensing
QI        Quality Improvement
RITS      RSS Inventory Tracking System
RSS       Receipt, Staging, and Storage
SNS       Strategic National Stockpile
TAR       Technical Assistance Review
ACKNOWLEDGMENTS

We gratefully acknowledge the large number of individuals who provided insight, guidance, and criticism at all stages of this project.

We begin by thanking the staff of the Centers for Disease Control and Prevention’s Division of the Strategic National Stockpile (CDC/DSNS), especially Stephanie Dulin, Linda Neff, and Pamela Nonnenmacher, for helping us to understand the operational realities faced by state and local health departments and for critiquing various portions of the report and operations manuals. We also thank Tim Quinn, Wrenn Slocum, Mike Moore, and Jason Zimmerman, who helped us during our early work with DSNS, upon which this report is built.

Dozens of state and local health department officials patiently tutored us on key aspects of the SNS process, allowed us to observe exercises or use their sites as field-test locations, and offered comments on and criticism of draft documents. Our efforts to engage practitioners were helped immeasurably by staff at the National Association of County and City Health Officials (NACCHO) and the Association of State and Territorial Health Officials (ASTHO)—especially Steve Curren, Jennifer Nieratko, and Katie Dunkel.

We also acknowledge the insights and assistance of RAND colleagues Karen Ricci, Stacey Rasfeld, and Michelle Horner on various aspects of the project, and Nicole Lurie and Jeffrey Wasserman for providing insight, support, and encouragement at all stages of the project.

Finally, we thank William F. Raub, Lara Lamprecht, and Matthew Minson at the U.S. Department of Health and Human Services Office of the Assistant Secretary for Preparedness and Response (HHS/ASPR) for their support and guidance throughout this project.

We alone, however, bear responsibility for the content of the report and the operations manuals.

Christopher Nelson
Edward Chan
Carol E. Fan
Debra Lotstein
Leah B. Caldarone
Shoshana Shelton
Amy Maletic
Andrew Parker
Alexandria Felton
Amanda Pomeroy
Elizabeth Sloss
CHAPTER 1: INTRODUCTION

State and local governments play the key role in receiving, distributing, and dispensing materiel from the Strategic National Stockpile (SNS). Assessing state and local departments’ readiness to perform these functions is a critical element for improving their ability to conduct a countermeasure-delivery campaign in an emergency.

The Centers for Disease Control and Prevention’s Division of the Strategic National Stockpile (CDC/DSNS) has been using a checklist-based approach, known as the Technical Assistance Review (TAR), to assess the jurisdictions’ countermeasure-delivery plans, equipment, and personnel. Summary results of these assessments have been published by CDC\(^2\) and by organizations such as the Trust for America’s Health.\(^3\) The results indicate that all 50 states (as well as the District of Columbia) have adequate plans for distributing countermeasures from the SNS. However, a consensus has emerged within the Department of Health and Human Services (HHS) and CDC that assessing these *capacity* elements is not sufficient; methods are needed for assessing jurisdictions’ *capability* to implement and adapt these plans in real-world situations.

Initially, the interest in measurement was for identifying areas in which state and local health departments need assistance to improve readiness. The need for measurement has increased with the passage of the Pandemic and All-Hazards Preparedness Act (PAHPA) of 2006 (P.L. 109-417), which, among other things, required HHS to develop evidence-based performance standards and metrics for preparedness, and, eventually to link decisions on funding for state and local agencies to their performance on these measures. The U.S. Department of Health and Human Services, Office of the Assistant Secretary for Preparedness and Response (HHS/ASPR) asked RAND to assist CDC/DSNS in developing methods for assessing jurisdictions’ capabilities to do countermeasure delivery.

This report presents a detailed summary of the assessments and the results of the field tests. The operations manuals that are discussed in this report provide guidance on assessing a jurisdiction’s countermeasure-delivery capabilities. Users can download and print spreadsheet-based data-reporting tools that will facilitate calculation of required metrics and collection of key data. The manuals and the data-reporting tools are available on the CDC website at: [http://emergency.cdc.gov/cotper/coopagreement/](http://emergency.cdc.gov/cotper/coopagreement/).

Throughout this report and the operations manuals, we use the term “patient” to refer to the individuals that visit PODs to receive countermeasures. We chose this term to be consistent with the language used in the POD modeling programs referenced in the Dispensing assessment manual. In an actual emergency, PODs would be expected to

---

\(^2\) CDC (2008).
\(^3\) Trust for America’s Health (2008).
serve healthy, non-injured persons, who would not be considered patients in the strict medical sense. True patients, i.e., sick or injured individuals, would go to healthcare facilities for treatment instead.

**ASSESSMENT DEVELOPMENT PROCESS WAS GUIDED BY EXPLICIT DESIGN CRITERIA**

The assessment development process was guided by the following principles, which were derived from discussions with CDC/DSNS, analysis of key SNS and Cities’ Readiness Initiative (CRI) documents, and earlier HHS/ASPR-sponsored RAND research on measuring Public Health Emergency Preparedness (PHEP) (e.g., Nelson, Shugarman, Lurie, and Hartman, 2006). In particular, the assessments should

- **Measure jurisdictions’ ability to implement and adapt preparedness plans.**
  Most important, the proposed system must assess jurisdictions’ ability to mobilize response infrastructure and resources and adapt response plans. Thus, the proposed system includes greater use of operational assessments (drills) than does the current system.

- **Impose reasonable assessment burdens.** The best assessment of operational capability is a documented response to a real-life emergency or, short of that, a full-scale exercise that simulates the scale and stresses of real emergencies. However, large-scale emergencies are (fortunately) rare, and full-scale exercises impose heavy burdens on assessed jurisdictions and the CDC. Thus, the proposed system is based on a set of small-scale drills that test smaller, common-mode components of a full-scale response, with larger exercises conducted only occasionally.

- **Support both accountability and quality improvement.** The SNS assessments must provide information to support both accountability decisions (which require standard metrics that can be compared across jurisdictions and over time) and process-improvement efforts (which requires more-customized data collection capable of supporting individual improvement efforts). Thus, we sought to develop standard—or semi-standard—metrics for each of the drills, but also designed opportunities to adapt and combine the drill formats to fit a wide variety of state and local improvement needs.

- **Cover a broad spectrum of jurisdictions.** The current expansion of the Cities Readiness Initiative program will require assessment of a larger number of jurisdictions (and with more-variable assessment capacity) than in the past, and some of these jurisdictions might have relatively little experience running drills and exercises. Thus, we sought to design assessments that could (if necessary) be reliably self-administered by a broad range of jurisdictions, with objective, quantitative, metrics that can be collected by jurisdictions without requiring high levels of training or expert judgment. Trained, outside evaluators may be used (perhaps from neighboring jurisdictions); however they are not strictly necessary. Additionally, we designed the assessments with as much flexibility as possible, so
that they can be readily adapted to different jurisdictions’ countermeasure-delivery plans, rather than forcing jurisdictions to conduct their operations in a particular manner or using a particular technology.

- **Be aligned with relevant guidance and doctrine.** Assessments will be more efficient and less frustrating to state and local jurisdictions if they are systematically aligned with relevant program guidance. Thus, we sought to align our efforts with CDC/DSNS’s *Receiving, Distributing, and Dispensing Strategic National Stockpile Assets* (version 10.02) and the U.S. Department of Homeland Security’s *Target Capabilities List* (https://www.llis.dhs.gov/displayContent?contentID=26724). At a more conceptual level, the proposed assessments are grounded in the capabilities-based approach to emergency planning that underlies Homeland Security Presidential Directive 8 (HSPD-8)⁴ and the modular and scalable approach to emergency response articulated in the National Incident Management System (NIMS) (http://www.fema.gov/pdf/emergency/nims/NIMS_core.pdf).

**EACH CHAPTER FOCUSES ON A KEY QUESTION**

The remainder of this report is organized around answers to four important questions:

- **Which SNS elements should be assessed?** Chapter 2 presents a model for Public Health Preparedness and Response that is used to generate a short list of SNS elements around which to develop assessments. Details on the analytic process used to select SNS elements are provided in Appendix A.

- **How should those elements be assessed?** Chapter 3 describes the main features of the proposed assessments and provides an overview of each of the operations manuals. The complete operations manuals for each of the proposed assessments are available on the CDC website at: http://emergency.cdc.gov/cotper/coopagreement/. On this site, users can also download and print spreadsheet-based data-reporting tools that will facilitate calculation of required metrics and collection of key data.

- **How well did the assessments work during field tests?** Chapter 4 provides a description of the methods used to conduct field tests and summarizes key findings from the field tests.

- **What are the next steps in developing and rolling out the proposed assessments?** Chapter 5 lays out recommendations for providing the assessments to jurisdictions, setting performance standards, linking assessments with state/local quality-improvement efforts, attaching consequences to assessment scores, and improving the evidence base behind assessments and standards.

---

CHAPTER 2: DETERMINING WHAT TO ASSESS

RAND’s charge was to develop assessments that test operational capabilities (i.e., the ability to actually implement key SNS capabilities and functions) but do not impose excessive burdens on assessed jurisdictions or CDC/DSNS. Keeping the operational assessments focused on a relatively small number of critical SNS elements is essential to simultaneously satisfying both of these criteria.

There is little evidence base regarding which aspects of preparedness are most important to measure. What evidence base there is for preparedness processes comes from after-action reports. However, large-scale public health emergencies are rare. Further, variation in what data is collected from events or from exercises makes it difficult to compare performance across different events, much less link performance to health outcomes.

In the absence of adequate empirical data, we employ a process engineering–based approach that seeks to identify core “building block” capabilities that can be deployed, combined, and adapted to a broad spectrum of response scenarios. The approach complements the capabilities-based planning (CBP) philosophy articulated in Homeland Security Presidential Directive 8 (HSPD-8) (http://www.dhs.gov/xabout/laws/ge_1215444247124.shtm), the Homeland Security Exercise Evaluation Program (HSEEP) (https://hseep.dhs.gov/pages/1001_HSEEP7.aspx) and elsewhere. The CBP approach is well-suited for situations such as public health emergency preparedness, in which there is considerable uncertainty about the requirements of unforeseen emergencies and disasters.

This chapter describes the capabilities-based approach, it discusses the candidate list of capabilities identified for assessment and approaches for administering assessments.

---

5 The capabilities-based approach to planning was developed by the U.S. national security community in response to the post–Cold War challenge of developing and maintaining readiness for an increasingly complex international security environment and increases in asymmetric threats (see, e.g., Davis, 2002a).

6 The problem of developing robust plans in the face of uncertainty is extremely prevalent in human life. For instance, in preparing for a lunch rush, the short-order cook must ready materials and staff without knowing exactly which plates will be ordered, and in what quantity. As does the capabilities-based planner, the cook might prepare ingredients that appear in a wide variety of menu items, leaving until the last minute decisions about how to combine these ingredients in response to specific orders. The problem is also well-summarized in a quote attributed to former NASA Flight Director Gene Krantz, in the film Apollo 13. Krantz is shown to remark, “I don’t care what the thing was designed to do; I want to know what it can do.” Here, again, the art of response (which must be addressed in preparedness for responding) lies in the ability to adapt previously developed and planned procedures to meet rapidly unfolding situations.
ASSESSMENTS BASED ON THE CAPABILITIES-BASED APPROACH

Public health emergency preparedness can be viewed as comprising two components. The *planned* component relies on using plans, training, exercises, leadership, and other processes to develop equipment and personnel into capacities that can be deployed in response to a broad spectrum of potential threats. The *adaptive* component involves the “improvisational” capability to assess a situation and deploy, adapt, and combine building-block capabilities that can reduce morbidity and mortality.7

Figure 2.1 represents a capabilities-based understanding of preparedness (including, but not limited to, SNS preparedness) in terms of the familiar input-process-output-outcome framework common in the literature on program design and evaluation. Inputs and processes constitute the static portion of the model. *Inputs* (shown in the leftmost box) are the “raw materials” of preparedness and include personnel and equipment. These inputs, in turn, are transformed via preparedness *processes/activities* (e.g., planning, training, exercises, organization/leadership) into *outputs*, which we conceptualize as the building-block capabilities described above8 and which make up the dynamic component preparedness.

These building-block capabilities (shown as separate boxes) are deployed and assembled (represented in the diagram as a set of connected blocks) in response to emergencies. If the match between capabilities and ongoing situational requirements is sufficient, the *outcome* of the response (shown at the right-hand side of the diagram) will be a reduction in morbidity and mortality. This match is ensured by appropriate detection, reporting, and investigation, along with sound decisionmaking about which combination of capabilities is most likely to reduce mortality and morbidity.9

7 Writing of defense planning, Davis (2002b) observes that, “success . . . depends on the suitability of the building blocks and the organization’s prowess in quickly assembling and controlling their integrated application to missions.”

8 Taken together, the inputs and processes comprise what DHS (and before that the Federal Emergency Management Agency [FEMA]) documents refer to these as the key “capability elements” (see DHS, 2005; FEMA, 1996).

9 These components are left out of the diagram for simplicity’s sake.
The rarity of large-scale public health emergencies, while fortunate, means that there is little empirical evidence on which to base identification of the most critical drivers of, or barriers to, successful SNS responses. Thus, we have relied on the conceptual model of preparedness as well as on engineering-style process analysis to identify a short list of SNS capabilities that seem important enough to merit assessment.

We used critical-path analysis (Kelly, 1961; McClain, Thomas, and Mazzola, 1992) to identify the “rate-limiting factors” in the SNS process--those things that, if not done in a timely fashion, could prevent delivery of medications to the public within 48 hours, as required by the Cities Readiness Initiative guidance (CDC, 2006). Examination of these factors helped identify capabilities most in need of assessment. Appendix A provides details on this analysis.

Figure 2.2 shows a set of capabilities identified in the critical-path analysis that might be candidates for assessment. The set of candidate capabilities are those most directly involved in the movement of materiel and people and is not intended to be exhaustive (e.g., it does not include capabilities related to epidemiology, surveillance, lab testing). The capabilities can be grouped into:

- Command and management capabilities: are for decision-making at a more strategic level, as compared to the more tactical decisions that would be made at particular facilities. In many departments, these capabilities would be included in the functions of the department’s operations center.
- Function-specific capabilities: are unique to specific functions or facilities in the countermeasure-delivery process, such as distribution; receiving, storing, and staging (RSS) materiel from the SNS; and mass dispensing.
- Crosscutting capabilities: are those activities that are common to a variety of tasks; for example, the capability to notify and mobilize staff (staff call-down) is
necessary for all SNS tasks that require additional personnel such as activating the Emergency Operations Center (EOC), the RSS warehouse, or point of dispensing (POD) site.

Assessment tools were ultimately created for five building-block capabilities. Two focus on functions specific to mass countermeasure delivery: dispensing and pick-list generation (part of RSS operations). Two focus on crosscutting capabilities: staff call-down (which includes hypothetical assembly) and site call-down. The fifth—site setup—is, in principle, a crosscutting capability but will look quite different depending on the functional context in which it is implemented; for this report we focused on POD setup.

In Figure 2.2 below, the selected capabilities are represented as Lego®-like building blocks and are color-coded according to capability area.

![Figure 2.2: Capabilities Selected for Assessment Tool Development](image)

Two considerations guided our decision to develop assessments for these five capabilities.

Focus on movement of materiel and staff. First, we decided to focus on capabilities that involve the movement of goods and the assembly of staff. We recognize the importance of assessments that test decisionmakers’ ability to determine when to activate procedures (e.g., requesting SNS), allocate resources (e.g., apportionment of materiel across Points of Dispensing), and so on. However, developing assessments for these capabilities
requires measurement of highly qualitative and context-specific skills, and therefore was judged outside the scope of this current project.10

Focus on the most failure-prone capabilities. We also focused on capabilities that seemed most failure-prone. For instance, in the case of assessing capabilities associated with operating the RSS warehouse, conversations with subject-matter experts and our own observations of exercises suggested that unloading SNS shipments is a fairly straightforward process as long as an RSS layout plan exists, a suitable facility has been activated, and enough RSS staff can be called down for service.11 These aspects can be verified through other drills and through the checklist-based assessment tool. Similarly, it did not seem to make sense to hold drills of only the delivery process; if sufficient trucks and trained drivers are present with appropriate directions, it is reasonable to assume that drivers can drive the trucks to the assigned location.

ASSESSMENTS ARE DESIGNED TO BALANCE STANDARDIZATION AND FLEXIBILITY

The proposed assessments reflect the logic of the conceptual model of PHEP presented in Chapter 2. Just as the model envisions preparedness as a matter of developing, deploying, and combining building blocks, the assessments are based on a set of small-scale, standardized, modular, scalable, extensible, and embeddable assessments of key capabilities. Each of these system characteristics helps address the design requirements outlined in Chapter 1.

The fact that the drills are small in scale helps the assessments balance the need to test operational capabilities with the desire to limit burdens on assessed jurisdictions. For instance, the pick list generation drill—the RSS-relevant drill described below—tests jurisdictions’ ability to generate instructions to warehouse staff for picking and packing materiel from the warehouse for shipment to POD locations, hospitals, or other sites. Jurisdictions often have trouble with this step in their full-scale exercises, bringing operations to a screeching halt. A drill focused on that key step can be conducted with just a few staff and requires only a small room, a computer, and minimal office equipment.

For the drills to be useful, a standard set of performance metrics must be collected. The Exercise Evaluation Guides (EEGs) developed by the Department of Homeland Security’s (DHS) Homeland Security Exercise Evaluation Program (HSEEP) have been one effort to define a uniform set of data to be collected at exercises. However, because

---

10 Assessing crisis decisionmaking for public health emergencies was the subject of a separate study. See Parker et al. (2009).
11 An argument could be made that picking and palletizing material at the warehouse are failure-prone enough to warrant a separate assessment, particularly when these tasks are performed by volunteers for whom operating a warehouse is not their regular job. However, resource constraints forced us to limit the number of assessments included in this initial development process. Moreover, a picking-and-palletizing drill would require such a large amount of equipment and supplies that a jurisdiction may as well conduct a full RSS drill.
EEGs provide little guidance on the scenario, or how the exercise itself should be conducted, considerable variation can occur, making comparisons across exercises difficult. Furthermore, although the metrics that currently exist in the EEGs tend to focus on the correct performance of tasks, they do not tend to focus on speed and scale, making it difficult to estimate the jurisdiction’s ability to conduct a mass countermeasure-delivery campaign, for which speed and scale are of the essence.

Each drill, therefore, includes standard metrics and provides enough standardization in the assessed tasks to ensure at least a modicum of comparability across jurisdictions and over time. Most of the metrics are for the amount of time required to complete critical steps in the SNS process. But standard metrics alone are not enough to ensure comparability. Thus, the operations manuals for each drill include detailed instructions on how to structure the assessed task, equip evaluators, and collect and report data. However, the manuals stop well short of prescribing complete, detailed exercise formats, which helps maintain their applicability to a broad range of exercises.

The drills can be implemented as stand-alone assessments or combined with other drills or assessment activities to form compound assessments, as illustrated in Figure 2.3, which shows the results of “snapping” the Lego®-like blocks together. For instance, assessments of function-specific capabilities such as RSS/pick-list generation and POD dispensing require processes for notifying staff and setting up facilities. They could be readily combined with assessments of crosscutting capabilities, such as call-down, site activation, and facility setup. In the figure below, the selected capabilities are represented as Lego®-like building blocks and are color-coded according to capability area.

Figure 2.3: Crosscutting and Function-Specific Capabilities can be Combined into Compound Drills

12 Thus, for instance, staff notification and assembly times from high-tempo scenarios (e.g., anthrax) are not distinguished from slower-tempo scenarios, such as foodborne outbreaks. See Nelson et al. (2008) for further discussion of these issues.
Thus, the assessments can be adapted to meet state and local quality-improvement needs without sacrificing policymakers’ need for standardized—and therefore comparable—metrics that can provide an overall summary of performance. Because the standardized drill modules are fairly limited in scope, they need not dominate ongoing training and assessment schedules. The drills can also be expanded, allowing jurisdictions to add locally relevant metrics to the standard set.

It might also be possible to embed operational-preparedness assessments in more-routine public health functions, thus reducing the need to assemble key staff solely for exercising. For example, jurisdictions might use annual refresher training for POD staff as an opportunity to assess the entire call-down and assembly process. Instead of telling people simply to assemble at an appointed hour, jurisdictions could instruct residents to not leave their homes/offices until they received a call telling them to do so. Having gathered all applicable staff, a jurisdiction might take the opportunity to walk staff through the facility-setup process.

Several jurisdictions are already using nonroutine practices or small-scale emergencies to test large-scale response capabilities. For example, several jurisdictions used the influenza-vaccine shortfall of 2004 as an opportunity to test the public health department’s incident command structure during an effort to provide mass vaccination to a large number of individuals in a single day (Seid et al., 2006). Other jurisdictions have used annual flu clinics to test mass dispensing procedures for the SNS program, which alleviates some of the costs associated with recruiting volunteers to act as test patients. Similarly, a local jurisdiction in Virginia recently used a tuberculosis-screening process as an opportunity to exercise incident command structures (Rendin, Welch, and Kaplowitz, 2005).

ASSESSMENTS CAN IMPROVE INFORMATION GLEANED FROM EXISTING EXERCISES

Currently, CDC/DSNS administers SNS assessments to state health departments and CRI-related local health departments during site visits. As noted in Chapter 1, however, our assessment design process proceeded on the assumption that the range of assessed jurisdictions might expand considerably, in part as a result of the ongoing expansion of the CRI program.

Continuing to administer assessments during site visits would have clear advantages. For instance, instead of having jurisdictions self-report that a Memorandum of Understanding (MOU) exists, site visitors can inspect the document themselves and judge its adequacy.

---

13 During such visits to complete the checklist -- the CDC/DSNS rater meets with staff members of the health jurisdiction and uses multiple data sources -- including document review, interviews, and observation. Many of the checklist items focus on the specific content of the jurisdiction’s SNS plan. For these items, the rater either reviews the written SNS plan or relies on responses to specific questions during interviews with staff members.
Site visits might provide opportunities for evaluators to speak with jurisdiction staff to
gauge their awareness and understanding of preparedness plans and resources.

However, the fact that, according to CDC/DSNS, staffing constraints have reduced the
number of site visits from twice to once a year makes it unlikely that the agency could
continue site visit–based administration with more jurisdictions and with the addition of
the capability-based assessments described above. Thus, the system must be designed
around the assumption that jurisdictions will self-report most or all data.

The system is also designed to increase the quality of evaluative information that can be
extracted from the already-large number of SNS exercises conducted by jurisdictions.
While data on SNS-specific exercises is not available, the CDC reported nearly 2,000
PHEP-related exercises during fiscal year (FY) 2005 (CDC, 2008). Yet, variations in
exercise structure, reporting practice, and the format of after-action reports makes it
difficult to make comparisons over time or across jurisdictions (see Nelson, Beckjord,
Chan, et al., 2008). Using the assessments presented in this report can allow for a higher
degree of comparability between critical aspects of SNS exercises.
CHAPTER 3: OVERVIEW OF OPERATIONS MANUALS

To ensure consistent and easy use of the drills across jurisdictions and over time, we developed short operations manuals for the assessment tools. The manuals are designed to support self-assessment by most jurisdictions. Initial field testing has been conducted (see Chapter 4); however, further field testing will be required to ensure that the manuals are usable by a broad spectrum of jurisdictions.

The manuals were drafted by RAND staff after a review of the relevant literature, discussions with subject-matter experts and practitioners, and observation of exercises. Drafts were critiqued by CDC staff and a limited number of health department staff. Each of the manuals uses a common outline, summarized in Table 3.1.

The manuals are not intended to be complete exercise plans. Rather, they are to be used by jurisdictions in conjunction with their own exercises. The manuals serve as guidance on how to collect data to compute metrics on important aspects of the jurisdiction's capability to perform countermeasure delivery.

Table 3.1: Content of Operations Manuals

<table>
<thead>
<tr>
<th>Section of Manual</th>
<th>Description of Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>An overview of the drill, including a brief description of the capability to be assessed, summary of required metrics, and any special issues that bear emphasis.</td>
</tr>
<tr>
<td>Preparing for the drill</td>
<td>A list of key tasks that must be completed before the drill and data collection begin. Typically, this step involves recruiting staff and volunteers, and ensuring that facilities and equipment are available and that relevant plans and protocols are ready.</td>
</tr>
<tr>
<td>Data-collection during the drill</td>
<td>Detailed instructions for conducting the drill and collecting relevant data in a way that ensures a reasonable degree of consistency across jurisdictions and over time.</td>
</tr>
<tr>
<td>Computing metrics after the drill</td>
<td>Detailed instructions for how to use the data collected to calculate the metrics.</td>
</tr>
<tr>
<td>Data collection spreadsheet</td>
<td>A spreadsheet-based tool that helps users record data and calculate metrics.</td>
</tr>
</tbody>
</table>
SELECTION OF MEASURES

Exercises that are conducted by health departments are often used for teaching purposes, to introduce staff and volunteers to procedures they may have to perform in an emergency. The simple accomplishment of tasks is considered the metric of success. While this introductory purpose is necessary in the early stages of a jurisdiction’s preparedness efforts, it is not sufficient for assessing whether a jurisdiction can conduct mass countermeasure delivery at the speed and scale that would be necessary, particularly in a CRI-style anthrax scenario, in which an entire population of a jurisdiction would need countermeasures within 48 hours.

Most of the metrics described below assess the amount of time it takes to complete the activity. As discussed in the preceding chapter, activities involving the movement of materiel are critical because they ultimately affect the time in which the population receives their countermeasures. Timeliness is therefore of the essence. However, other activities that are prerequisites to the movement of materiel, such as mobilization of staff and set-up of facilities, must also be conducted in a timely fashion or else they will delay the ultimate movement of materiel.

In addition to speed, scale is also important. For dispensing, throughput metrics enable calculation of the overall rate at which the population would receive prophylaxis. For call-down, estimation metrics enable a count of the number of people who can be mobilized to staff the countermeasure-delivery operations.

In the interest of keeping burdens low, we limited the measures that would be collected. We were concerned that act of collecting one measure could affect another measure: In the call-down drill, for example, attempting to collect too much information on each call could distort the measure of the time necessary to make calls. We also sought to avoid metrics that would require high levels of judgment to measure. Consequently, we focused on quantitative rather than qualitative measures. Future measure-development activities should address the more-qualitative dimensions of preparedness and response. However, doing so will require additional research to identify the most critical quality dimensions.

In addition to collecting metric data, each assessment tool asks for a small amount of data on the conduct of the drill and on a jurisdiction’s approach to the assessed task. For instance, the tool for the call-down drill asks jurisdictions to provide information on communications equipment and protocols. All drills ask for basic staffing information. Such information might provide an important and convenient data source for examinations of whether certain ways of structuring responses are associated with better performance on key metrics, and thus helping to deepen elements of the PHEP evidence base.
The complete operations manuals for each of the proposed assessments are available on the CDC website at: [http://emergency.cdc.gov/cotper/coopagreement/](http://emergency.cdc.gov/cotper/coopagreement/). This chapter provides a concise overview of the current versions of the five assessment manuals and describes the main features of each (see Table 3.2). Readers interested in more detail should turn to the manuals themselves. Questions that were used to elicit feedback on field tests are included in Appendix B.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Overview of Assessment Task</th>
<th>Metrics (required and optional)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff Call-Down</td>
<td>Unannounced call-down drill; respondents indicate ability to report, but no actual staff movement required (hypothetical assembly).</td>
<td>Required:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Call-down completion time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Acknowledgment completion time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Acknowledgment percentage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o 50th percentile of acknowledgments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o 75th percentile of acknowledgments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Assembly percentage (hypothetical).</td>
</tr>
<tr>
<td>Site Call-Down</td>
<td>Unannounced calls placed to managers/owners of facilities used for SNS operations (e.g., warehouses, POD sites); respondents indicate availability of facility to be made ready for SNS use.</td>
<td>Required:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Site-call-down completion time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Acknowledgment completion time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Acknowledgment percentage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o 50th percentile of acknowledgments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o 75th percentile of acknowledgments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Availability percentage.</td>
</tr>
<tr>
<td>Pick-list generation (for RSS warehouse)</td>
<td>Generation of pick list in response to approved orders or Department Operations Center–approved apportionment strategy. May be conducted as a stand-alone drill or as part of a larger RSS drill.</td>
<td>Required:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Time to generate pick list</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optional:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Time to input inventory data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Time to add new item</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Time to add new destination.</td>
</tr>
<tr>
<td>POD Setup</td>
<td>Pre-announced physical setup of a POD facility.</td>
<td>Required:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Facility-setup completion time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Checklist completion rate.</td>
</tr>
<tr>
<td>Dispensing (POD)</td>
<td>Study to measure processing time for each step in a POD. Data from time study are input into a computer model to estimate maximum possible throughput of the POD.</td>
<td>Required:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Patient throughput (per hour).</td>
</tr>
</tbody>
</table>
OPERATIONAL ASSESSMENTS

Staff Call-Down

*Staff call-down* is the ability to contact and mobilize staff to perform emergency-response functions. Call-down is a crosscutting capability, applicable to a wide variety of functions, including dispensing, warehousing, distribution, security, and command centers, among others.

Since the goal is to gather staff (who might be paid or volunteer) for duty in an emergency situation, the calls must be conducted in a timely fashion, and the calls must reach all the necessary personnel. The end result in an actual emergency should be that the jurisdiction can gather all the necessary personnel at the time they are required. Consequently, the call-down drill tests the validity of jurisdictions’ call-down lists and their ability of jurisdictions to contact those staff in a timely manner; it estimates the percentage of staff who could report for duty within a designated time frame.

An ideal call-down drill would be conducted without notice and would involve actual assembly at a site. However, requiring staff, especially volunteer staff, who normally have other commitments, to frequently assemble for drills without any advance warning could create undue burdens that ultimately reduce the willingness of volunteers to serve. Moreover, it is not clear that travel times derived from drills would reflect real emergency conditions (e.g., traffic, compromised infrastructure). Therefore, the procedures and metrics outlined are based on a no-notice, hypothetical assembly, call-down drill. That is, staff members are contacted using whatever communications system would be used during a real emergency (e.g., automated calling system, calling tree) and asked whether they could report to a given location at a certain time, but they are not actually required to assemble.

Data are generated from a no-notice call-down drill and require an evaluator(s) or a calling system to record, for each person contacted, the time he/she acknowledged the call and whether he/she could assemble by a stated target time.

Performance metrics for this assessment include:

- **Call-down completion time**: the amount of time required to call all staff on the call-down list
- **Acknowledgement completion time**: the amount of time required to receive response from staff confirming receipt of the call-down message, regardless of ability to assemble.
- **Acknowledgement percentage**: the proportion of staff on the call-down list who confirmed receipt of the call-down message, regardless of ability to assemble.
- **Assembly percentage**: the proportion of staff on the call-down list who reported being able to assemble at a designated site by a predetermined target time.

The completion-time metrics serve to assess the amount of time that would be needed to notify staff members. The acknowledgment completion time is needed in case messages were left. The acknowledgment-and-assembly percentages are useful for determining the proportion of personnel who would be unreachable or unavailable on a given day, providing an estimate of the extra personnel who would need to be recruited as backups.

Because this metric relies on the self-report on the part of those called, it could result in an overestimate of assembly percentage in which more people say that they are able to report, than would actually show up in an emergency. On the other hand, forcing an actual assembly could result in an underestimate of assembly percentage, as some people, knowing this were not a true emergency, might not show up for the drill due to the inconvenience. In any case, it is difficult to estimate a person’s willingness to serve in an emergency via any questions asked during a non-emergency situation. We considered adding more questions to be asked during the call-down drill, such as inquiring about the status of the called person’s family members, but concluded that this would lengthen the call and thus distort the measurement of the call-down completion time. We also considered asking those called to volunteer the time at which they thought they could assemble, rather than asking them whether they could report by a certain time. However, this would add to the data collection burden, particularly in the case of automated systems which might not be able to accommodate such open-ended responses.

**Site Call-Down**

In a public health emergency, health departments may need to make use of facilities whose everyday functions are not dedicated to countermeasure-delivery purposes, and thus whose availability may be in question. This questionableness is especially the case with, although not limited to, facilities that are not ordinarily under the health department’s ownership or control. Examples of such sites include school buildings or community centers that would serve as PODs, or storage warehouses that would serve as RSS warehouses. The site-call-down drill tests jurisdictions’ ability to quickly contact operators and/or owners of sites that would house critical SNS facilities, to determine how quickly the sites could be made ready for emergency operations.

Similar to the staff call-down, an ideal site-call-down drill would be conducted without notice and involve actual site preparation for setup. However, requiring sites, many of which normally function as schools, warehouses, or medical centers, to interrupt their normal non-emergency functions without warning might create an undue burden. While such tests should be conducted periodically, procedures in the manual are based on a no-notice, hypothetical site-activation drill to permit more-frequent, lower-cost assessments.

This assessment tests the operational capability to call sites when an emergency occurs. Performance metrics include
- **Site-call-down completion time**: the length of time required to call all sites on the site-call-down list.
- **Acknowledgment completion time**: the length of time required to receive response from sites confirming receipt of the call-down message, regardless of ability to make the site available.
- **Acknowledgment percentage**: the proportion of sites on the site-call-down list that confirmed receipt of the call-down message regardless of ability to make the site available.
- **Availability percentage**: the proportion of sites on the site-call-down list that reported being able to make their site available for use by the health department by a predetermined target time.

Just as with the staff call-down, this assessment requires self-reporting on the part of the site owners or operators who are called. It is possible that self-report would result in more people reporting that their sites could be made available than they actually would be in an emergency. However, there is also reason to believe that in an actual emergency, more heroic measures would be taken to make a site available. For example, interviews with site owners indicated that, in a true life-or-death situation, warehouse operators would sacrifice the materiel in their warehouse, bulldozing it out of the way if needed, so that the space could be used. However, they would be unwilling to take such drastic measures if the site were needed only for a drill.

### Pick-List Generation

Successful responses to emergencies require effective and timely matches between demand for and supply of critical resources. The assessment of pick-list generation tests jurisdictions’ ability to use their warehouse inventory system to ensure such a match. Operating the inventory management system is but one of many components of operating an RSS warehouse. However, it is akin to the command and control function for the facility. During exercises that we observed, difficulties in operating the inventory management system caused delays in picking material for shipment to PODs. During a real emergency, such problems would cause shortages at the PODs, resulting in an overall delay in the dispensing campaign. Since measuring all components of the inventory system would be time-consuming and difficult, this assessment focuses on one critical component identified through observation of drills and consultation with subject-matter experts: the generation of a set of warehouse pick lists. **Pick lists** are the instructions that tell workers at the warehouse which items should be retrieved and packaged for shipment to a location, such as a POD or a treatment center.

The assessment may be conducted as a stand-alone drill, in which jurisdictions do a run-through of the pick-list-generation process. Alternatively, it could be conducted as part of a larger RSS exercise without otherwise interfering with the RSS exercise. If
conducted as a stand-alone, all that is needed would be a room, computer equipment, and the relevant staff, making the exercise one that imposed few burdens on the jurisdiction. Because delays in generating the pick list were the sticking point in many of the exercises we observed, we constructed the assessment to require the following performance metric:

- **Pick-list-generation time**: the average number of minutes required for jurisdictions to generate a pick list for use by warehouse personnel. (Accuracy of the pick list is not explicitly measured but, rather, is assumed as a necessary condition of the proper generation of a pick list.)

Depending on whether data for the metrics are collected as part of a larger RSS exercise and what type of warehouse inventory system is used by the jurisdiction, the following optional performance metrics may also be computed:

- **Inventory data-input time**: the number of minutes required for jurisdictions to input inventory data into their warehouse inventory system.
- **Add-new-item time**: the number of minutes required for jurisdictions to add a new item to their warehouse inventory system.
- **Add-new-delivery-destination time**: the number of minutes required for jurisdictions to add a new delivery destination to their warehouse inventory system.

These additional metrics are intended to provide more detail about the time required for each of the steps in generating a pick list, as well as to encourage jurisdictions to practice more procedures on an inventory management system. Some of these procedures, however, may be incompatible with the overall plan of a jurisdiction’s RSS or full-scale exercise; for example, their exercise might not necessarily involve the sudden addition of a new delivery destination. Therefore, we made these additional metrics optional.

**POD Facility Setup**

The goal of a facility-setup drill is to test the number of hours required to completely set up a facility with the materiel, layout, and supplies necessary to perform a given SNS function. Facility setup is a crosscutting capability, applicable to a wide variety of SNS functions, including dispensing (POD), warehousing (RSS), and command and control (Emergency Operations Center [EOC]), among others. The notion of timing the set-up of a facility can be applied to a variety of facilities; however, the details of the setup are specific to the type of facility. In this report, we present an assessment that tests the ability to quickly set up a facility for use as a POD.

Data for the metrics come from timing staff as they set up the facility, and a walk-through after the setup to assess the quality of the setup and identify areas of potential improvement. The assessment may be conducted as either a small, stand-alone drill or by timing the setup process prior to using the POD for a larger dispensing exercise. Performance metrics include
- **Setup time**: the number of hours required to fully set up the POD.
- **Quality of setup**: the success in completing items on a checklist of key tasks.

Timely setup is an important precondition of a rapid and effective response and may, in some situations, lie on the critical path to successfully delivering countermeasures to a population within 48 hours. Consequently, the time needed to perform the setup is the main metric. However, for the assessment to be meaningful, there needed to be at least some general guidelines on what would constitute a properly setup POD. Since it would be impossible to generate a list that would at once be comprehensive and complete, and yet flexible to the variety of POD designs that exist, the check of setup quality does depend on self-policing on the part of those conducting and evaluating the exercise. The walk-through that is presented in the manuals also serves as a tool for assisting jurisdictions in identifying gaps in their POD designs.

**Dispensing**

This assessment tests the ability to quickly deliver countermeasures to large numbers of people at a POD. The assessment is designed to be used during a POD drill and focuses on measuring maximum possible throughput (i.e., the number of patients processed per hour). Throughput was deemed the essential metric, since the goal of the mass countermeasure-delivery campaign is to dispense the materiel to all the people who need it within the targeted time line. By assessing the rate at which countermeasures are dispensed under the jurisdiction’s POD design and staffing plan, jurisdictions can validate the estimates of the total number of PODs needed and the total number of staff needed for the countermeasure-delivery campaign.

Conversations with subject matter experts, observations of drills, and examination of after-action reports (AARs) from drills all suggest that jurisdictions often have a difficult time recruiting enough volunteers to stress the POD. Often, POD drills are run with 100 or 200 persons spread out over several hours, never creating an opportunity to test the POD design and staff’s ability to handle large numbers of people arriving at a POD at one time. This lack reduces the reliability of drill-based throughput estimates. Thus, we recommend a two-step approach:

- During the POD drill, conduct a **time study** to measure the processing time (i.e., how long each step takes) for each station. Evaluators posted at each station of the POD record the start time and end time for each patient observed. This time study can be performed without interfering in the operations of the POD.
- After the POD drill, input the processing-time information from the time study into a computer model to estimate the maximum possible **throughput** of the POD.

During the field testing of this assessment (described in the next chapter), concerns were raised by health departments regarding their familiarity and comfort with the use of computer models. In response to this concern, we developed a sixth assessment operations manual, a variation of the dispensing assessment that does not require the use
of computer models but, instead, relies on direct observation of the throughput of the POD, to be used when a sufficient number of volunteer patients can be recruited for the POD drill.
CHAPTER 4: FIELD TESTING METHODS AND RESULTS

The assessments were field-tested to obtain a preliminary assessment of their feasibility and utility, and to identify and address obvious deficiencies in the validity and reliability of the tests. The manuals were then revised based on feedback from the field testing, as well as from conference calls held with measurement and SNS workgroup members from the Association of State and Territorial Health Officials (ASTHO) and the National Association of City and County Health Officials (NACCHO).¹⁴

This chapter describes the methods and the results of field tests of four of the five assessment manuals: Staff Call-Down, Pick-List Generation, Dispensing Using Time Studies and Computer Models,¹⁵ and POD Setup. Because the Site Call-Down drill is nearly identical to the Staff Call-Down drill (differing only in the list being called and the questions asked during the call), it was not tested separately.

Data quality from the test sites varied widely; therefore, this report provides only general findings from the field tests. A more-extensive and systematic pilot-testing process will be required to establish with greater certainty the extent to which the assessments provide unbiased and low-variance indicators of state and local health departments’ SNS readiness.

FIELD-TESTING PROCEDURES

In the field tests, a small number of health departments used the assessment manuals to collect data from drills. In some cases, the departments conducted assessments entirely on their own; in other cases, RAND staff members performed the role of evaluator and assisted in the data-collection process. In addition, to accommodate site-level needs, we had to be somewhat flexible about which parts of the assessments each site tested. However, we are unable to provide any conclusions about the measurement characteristics or uses of the assessments for multiple and single components. We should also note that, in the pilot-test phase, by agreement with CDC/DSNS, there were no plans to conduct field tests for POD setup. However, an opportunity arose during the field-test phase to assess a POD facility setup, and we have included the results in this report.

¹⁴ For instance, even before pilot testing, stakeholders expressed concerns about the feasibility of performing call-down drills on POD volunteer lists in very large communities, which might involve calling (and maintaining lists of) thousands of individuals. In response to this concern, we developed a sampling strategy for call-down lists of greater than 500 people. The new sampling strategy is now included in the staff call-down assessment manual.

¹⁵ Based on concerns regarding health departments’ unfamiliarity with using computer models, we created a sixth assessment operations manual, a variation on the Dispensing assessment that does not involve the use of computer modeling. This assessment, “Assessing Mass-Dispensing Capability Using Direct Observation,” is included in Appendix B, but we did not have an opportunity to field-test it.
The level of RAND involvement was determined through negotiations with each site and generally reflected each site’s familiarity with exercising the capabilities reflected in the assessments. For instance, many health departments already conduct regular staff call-down drills; thus, it is not surprising that all the sites involved in field tests of this assessment were willing to use the materials on their own. Because the data-collection and computation requirements of the computer-based mass-dispensing assessment were less familiar to most sites, RAND played a more active role in these field tests.

Another consideration driving the level of RAND involvement was the extent to which primary drill activities could readily be observed. The activities of a staff call-down drill, for instance, are widely dispersed and difficult to observe, whereas activities in a dispensing drill are more centrally located and easier to observe. The list of drills, number and type of departments participating, and the level of involvement of RAND staff during the drill are shown in Table 4.1.
Table 4.1: Field Tests Conducted

<table>
<thead>
<tr>
<th>Drill</th>
<th>Number &amp; Type of Pilot Sites</th>
<th>Level of RAND Involvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff call-down</td>
<td>3 - State 3 - Local a</td>
<td>Departments conducted drills entirely on their own.</td>
</tr>
<tr>
<td>Site call-down</td>
<td>Not tested</td>
<td>Two of the field tests were conducted as stand-alone drills; RAND staff served as controller and evaluator for these drills.</td>
</tr>
<tr>
<td>Pick list generation</td>
<td>2 - State 1 - Local</td>
<td>One test was conducted in conjunction with an existing RSS exercise; RAND staff served as evaluator for the pick-list portion.</td>
</tr>
<tr>
<td>POD facility setup</td>
<td>1 - Local</td>
<td>Department conducted drill entirely on its own.</td>
</tr>
<tr>
<td>Dispensing using time studies</td>
<td>4 - Local</td>
<td>Departments planned and controlled all drills. RAND conducted the post-drill analysis for all dispensing drills.</td>
</tr>
<tr>
<td>computer models</td>
<td></td>
<td>Evaluators (timers) for the 4 field tests:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 1 - RAND staff served as timers</td>
</tr>
<tr>
<td>Dispensing using direct</td>
<td>Not tested</td>
<td>• 1 - RAND staff trained local timers and shared evaluator duties with them</td>
</tr>
<tr>
<td>observation</td>
<td></td>
<td>• 2 - Departments conducted timing on their own.</td>
</tr>
</tbody>
</table>

a The table below provides additional detail on the number of individuals called for each drill.

<table>
<thead>
<tr>
<th>Site</th>
<th>Characteristics of call-down drill</th>
</tr>
</thead>
<tbody>
<tr>
<td>County health dept.</td>
<td>1 - manual calling with calling tree (101 people)</td>
</tr>
<tr>
<td>County health dept.</td>
<td>2 - manual calling (&lt;10 people each)</td>
</tr>
<tr>
<td>County health dept.</td>
<td>1 - automated calling (149 people)</td>
</tr>
<tr>
<td></td>
<td>1 - manual calling (60 people)</td>
</tr>
<tr>
<td>State health dept.</td>
<td>1 - manual calling with calling tree (27 people)</td>
</tr>
<tr>
<td>State health dept.</td>
<td>1 - automated calling (55 people)</td>
</tr>
<tr>
<td>State health dept.</td>
<td>3 - automated calling (11, 31, and 270 people)</td>
</tr>
</tbody>
</table>
Regardless of the level of RAND involvement, health departments were given copies of the manuals ahead of time and received a telephone orientation to the assessment’s key requirements and procedures. In nearly all cases, the field test of the assessment was done in conjunction with already-scheduled drills. Thus, only departments that planned on conducting a drill expressed interest in pilot-testing the assessments, which limited the number of field tests that could be conducted, in turn limiting the generalizability of findings from the field tests. However, this compromise was necessary to keep the burdens on participating sites reasonable.

**Sampling**

We recruited state and local health departments to participate in the field testing of the assessments. Potential participants were identified through discussions with CDC/DSNS program service consultants. We also held conference calls with members of the SNS and measurement workgroups of ASTHO and NACCHO to solicit feedback on the manuals and possible test sites. Departments participated on a voluntary basis and were promised that their identities and performance in the assessment would not be published. We interviewed one to three staff members at each site. Most of the interviewees were exercise coordinators, but we also interviewed key players involved in the particular capability tested by the assessment (e.g., inventory control officer or automated calling system expert).

**Data Collection and Analysis**

Following the pilot tests, RAND staff used a semi-structured interview protocol (see Appendix B) to guide discussions with members of the health department about the assessment tools. The questions focused around four main topics:

- **Feasibility**: How much effort was involved in using the assessment? Was the burden reasonable?
- **Reliability**: Were there procedures or instructions that were confusing? Would a different set of evaluators using these same manuals record different results on the metrics?
- **Validity**: Do the metrics provide a fair and accurate picture of the capabilities? How well do they comport with any previous assessments that were done?
- **Utility**: Are the metrics useful as a tool for internal quality-improvement efforts?

As noted above, in some instances the pilot tests were undertaken without direct RAND involvement; for these, we conducted the interviews by telephone. We conducted in-person interviews at those sites at which we played a direct role. When possible, a lead interviewer and note-taker were included in the interviews. Interview data were coded thematically by the lead interviewer and validated by at least one other researcher directly involved in the interviews. The coding was then reviewed by one other researcher.
involved in the project but not directly involved in the interviews. Codes were derived inductively from the notes, but informed by the categories in the interview protocol.

In several instances difficulties in scheduling interviews meant that interviews were conducted largely impromptu, reducing the team’s ability to collect and analyze systematic data from those sites.

In addition to the qualitative data, pilot tests of the mass-dispensing assessment yielded data on station cycle times and, in some cases, on observed patient throughputs. These data were analyzed to produce estimates of mean and variance, and were used in conjunction with computer models to estimate maximum feasible throughput. All such data were shared with the pilot sites.

**Findings**
The remainder of this chapter summarizes key findings from the field testing.

**Assessments Were Viewed as Useful and Feasible**
The field testing indicated that, in general, the assessments are useful and feasible. The drills can be conducted and data collected without much burden. Field-test sites reported that the manuals and spreadsheet-based tools were very useful for managing data and calculating metrics in a standardized way. They also reported that having standard performance metrics provided useful feedback on the site’s performance in the drill. One respondent noted the value of having a “hard number on a page” to provide clear performance feedback.

**There Were Variations in the Amount of Technical Assistance Required**
We found that field-test sites needed different levels of technical assistance to conduct the drills and gather data. These levels varied by the site’s apparent skill and experience with preparedness exercises in general, and with the specific capabilities tested by the assessments.

At one end of the continuum, the call-down assessment required relatively little explanation, in large part because departments were accustomed to conducting call-down drills. As a result, sites were familiar with most of the procedures outlined in the staff call-down assessment manual. For the most part, the manual worked well with both automated and manual calling systems, and several sites reported that the data-collection sheet decreased the burden of their current drill procedures by assisting in calculating performance measures.

However, there was some concern that the acknowledgment and assembly metrics in the call-down manual are not feasible for many automated calling systems. For example, the assessment tool requires a record of whether the call recipient acknowledged the call.
Yet, many systems record the call as successful if a message is left, regardless of whether the call recipient actually received the message. Another challenge is the requirement to measure the percentage of call recipients who report being able to assemble for duty. Requiring these metrics might help spur demand for automated calling systems with this capability, but it is not clear whether manufacturers would respond to the demand in a timely manner.

As with the call-down assessment, the sites conducting the POD setup assessment reported that testing this capability was a familiar task. However, typically the POD setup is performed to acquaint staff with the layout of the POD, not as a test of their setup capability, and thus it is not usually timed. Nonetheless, interviewees indicated that conducting the assessment presented no significant difficulties.\(^\text{16}\)

The pick-list-generation assessment required more assistance and explanation, largely because the field-test sites were unfamiliar with the concept of a stand-alone drill devoted solely to exercising the pick-list-generation function. Thus, as noted above, RAND staff served as the controller and evaluator for two stand-alone field tests and evaluator for another instance in which the pick-list-generation drill was included in a larger RSS exercise. Consequently, we were unable to completely verify whether departments could conduct the pick-list-generation drill on their own.\(^\text{17}\)

**Using Time Studies and Computer Models Presented Feasibility Challenges**

As noted in Chapter 3, the computer-based mass-dispensing assessment requires two steps:

- Performing *time studies* of POD stations to measure patient processing time
- Running a *computer model*, using the time-study data, to obtain an estimate of maximum feasible throughput.

We discuss each in turn.

**Time Study.** As noted in Table 4.1, two of the field-test sites for the computer-based mass dispensing assessment conducted the timing on their own, and a third shared timing tasks with RAND staff. The field tests suggest that performing time studies is feasible for sites. However, sites expressed concern about the number of staff needed to do the timing. Of the two sites conducting the timing without RAND involvement, one used five staff and the other four. This does not seem like a high number of evaluators; however, both sites noted that exercise evaluators are often in short supply and were

---

\(^{16}\) One pilot site had to abandon plans to conduct the POD Setup assessment on a weekend day when too few workers showed up. However, this did not appear to be a result of the assessment itself.

\(^{17}\) None of the field-test sites used the CDC’s RSS Inventory Tracking System (RITS); all used other computerized tracking systems. However, it seems likely that the assessment would work equally well with RITS.
concerned that dedicating so many staff to this one aspect of evaluation (i.e., timing) might detract from other evaluation functions. Seeking to minimize the burdens of timing on POD staff, a third site asked their volunteer patients to record their own time stamps—the time, to the minute, at which some activity occurs. However, many patients failed to diligently record their time stamps, which limited the amount of data that could be collected.

**Computer Modeling.** Use of computer models was more problematic. The model suggested for use -- the University of Maryland’s Clinic Generator--is widely available, free of charge, and designed for use by a general audience; nevertheless, the field-test sites were generally not familiar with the concept of computer modeling or the mechanics of running the models. In addition, some staff reported that, given the pressure they were already under to prepare for their POD exercises, they were unable to devote much time to learning about computer modeling before the drill. (After the drill was completed, and with assistance of RAND staff, the pilot-site staff were able to become familiar with the software in only a couple of hours.)

That said, there is evidence that other sites might be more comfortable using computer models. Two other big-city health departments that RAND talked with as potential field test sites were familiar with computer modeling, and used models even more-sophisticated than the one we suggested.18

**Flexibility Built into Assessments Might Limit Data Comparability Across Sites**

Given considerable variations in public health systems and practices, the assessments were designed with an eye toward balancing the need for standardized and comparable data elements against the need to provide flexibility to sites. As a result, the assessments have up to this point provided considerable flexibility to sites in specifying exercise scenarios and response processes. However, the field tests made it clear that such flexibility might significantly limit the ability to draw valid comparisons among performance-metric data across sites and over time.

**Staff Call-Down.** As noted above, the process of testing staff call-down was quite familiar to each of the pilot sites and posed relatively few difficulties. Nonetheless, we observed a number of factors that could threaten the comparability of performance data derived from this assessment. Variations among field-test sites of the staff-call-down drill included the following:

- **Type and number of staff called.** The call-down assessment is intentionally agnostic on the type of staff to be called (e.g., operations center versus POD volunteers). Contacting volunteers will likely be more difficult than contacting

---

18 One possibility would be for sites to submit their data to a CDC/DSNS Program Services Consultant, who would, in turn, run the model. This adds to the burden placed onto CDC; it is not clear whether CDC would have sufficient staff to run the model for every site.
paid staff, who are more accustomed to being called. Similarly, the assessment
does not specify the number of staff to be notified in a call-down drill, which also
affects the degree of difficulty of the assessment. Decisions regarding the scale
and difficulty of the drill should be made by policymakers and standard-setters
based on their need to ensure readiness. For fieldtesting, we were primarily
interested in testing the mechanics of the drill, so we left these decisions to the
agency conducting the drill.

- **Time of day, extent of prior notice, and assembly requirements.** The pilot sites
varied by whether they conducted staff-call-down drills during normal business
hours or after hours and on weekends, as well as in the extent to which staff were
given prior notice of the drill. In addition, although physical assembly of staff is
not required by the assessment, some sites included it as part of the call-down
drills used to generate data for the assessment. Since these factors will likely
affect the degree of difficulty in contacting staff, comparisons of call-down
metrics across sites and over time might reflect differences in the conditions under
which specific call-down drills were conducted, as well as sites’ underlying
capability to perform this function.

- **Calling-tree procedures for reporting results.** We also found variations among
field-test sites in calling tree procedures for reporting results to a central location.
In some cases, individuals on the “branches” of the calling tree were instructed to
report call-down information immediately to a central point; in other cases, they
were allowed to wait until the next day to report their results to the central
authority. In the latter cases, call-down times will not reflect real-time emergency
capability, and variations in the call-down assessments might reflect site-by-site
idiosyncrasies as well as variations in underlying capability.

- **Script for outgoing messages.** Some sites did not follow the script for the
outgoing message provided in the call-down manual, which led to confusion
about the instructions given to call recipients, and potentially affected the
response rate. For example, some call recipients did not realize that a response
was required or did not know how to respond, because these directions were not
mentioned until the end of the call-down script.

**Pick-List Generation.** The primary challenge to the comparability of data in this
assessment is the variation in the sites’ warehouse inventory systems. Some warehouse
inventory systems are paper-based, others are Excel-based, and still others were created
specifically for the department. In addition, there are different policies for moving
materiel from the RSS to PODs and treatment centers. Some departments rely on a
“pull” system: The RSS sends materiel to PODs only in response to approved orders.
Other departments use a “push” system: The RSS predetermines the orders that each

---

19 Decisions about normal business as opposed to other hours and about degree of notice were often
influenced by concerns about staff burdens and labor contracts.
POD will receive. These two different policies can result in different pick-list-generation times.

In addition, the level of complexity in the pick list varies with the scenario and number of PODs. Some scenarios may require distribution of more types of materiel to more locations than others, making comparability across sites difficult.

**Mass Dispensing.** As noted in this chapter, a key component of the mass-dispensing assessment is conducting time studies at each station in the POD. From the field-test experience, it appears that evaluators across the sites used comparable measurement procedures. However, there appear to be differences in measurement among POD workers, and differences in the pace at which workers completed mass-dispensing tasks. The relatively small number of observations prevents us from conducting a rigorous statistical analysis of this effect. Exercise evaluators are therefore advised to time different workers within each station to obtain a more representative sampling of each station’s average processing time.

Much less reliable were time-stamping methods that depended on patients recording their own start and end times. Although, in theory, this method collects timing data for every patient going through the POD, the volunteer patients in the field test discussed above often forgot to record one or both of the times at each station or incorrectly attributed their times to the wrong station.

A final source of concern about the comparability of data from the mass-dispensing assessment lies in the mix of patients used during drills. Mass-dispensing field-test sites assigned volunteers roles that represented differing medical histories and conditions. Ideally, the mix of patients used in the exercise would reflect the mix of patients they would see in a real emergency. However, departments seeking to demonstrate a high throughput could potentially game the system by including only “easy” patients.

**POD Setup.** The main challenge in the reliability of the POD-setup assessment is determining the start time and stop time of the set-up process, and in the judgment of what constitutes a fully set up POD. Conceivably, the start time could be when workers are told to arrive at the POD, or when the full complement of POD workers have arrived and are ready to begin setup. Choosing the former would confound setup time with worker assembly time, so we opted for the latter. Defining when POD workers are ready to begin setup does, however, depend on agreement between the controller of the exercise and the POD manager. Similarly, it seems reasonable to designate the end time to be the time at which the POD manager declares to the exercise controller that the POD setup is complete. This seemed fairer than imposing a time deadline.

In an effort to impose some conditions on what constitutes a properly set up POD, we added one more performance metric to assess quality of setup—specifically, the percentage of POD-setup checklist items observed to be complete during a walk-through
following the POD setup. Even so, what constitutes a properly set up POD will vary with individual jurisdiction’s POD designs: Some POD designs will be larger or more complicated than others. There will also be variation in what is considered an acceptable departure from the stated POD design: Judgment is required in determining whether a missing sign or table should be considered as important as a missing rope line or computer.

**More Testing Is Required to Assess Validity**

Along with data comparability, a key question in the development of any assessment is the extent to which the assessment actually represents the capability it seeks to measure--i.e., to determine the validity of the assessment. It was not possible in this study to rigorously compare findings from the field tests with external benchmarks to help assess the degree to which the assessments are properly calibrated. However, in most instances, interviewees reported that findings from the assessment comported with findings from other drills and, in general, appeared plausible. Whether or not this provides evidence for validity, of course, depends on the quality of the prior assessments. Nonetheless, it is perhaps instructive that users found the results of the assessments to be plausible and in line with other assessments.

In particular, the computer-based mass-dispensing assessment presents some potential validity issues. It is clear that computer-model outputs are likely to overestimate the maximum possible throughput, because the models may not take into account real-life factors: e.g., errors by workers or patients, the delays between the end of processing one patient and the beginning of another (which, although seemingly slight, can add up), or other changes in performance caused by learning, stress, or fatigue. Thus, the results from the model should be used as a guide, not as a guarantee. As this method is implemented in the field, data from the dispensing drills--both the directly observed throughput and the estimates from the computer models--should be collected and compared.
CHAPTER 5: IMPLICATIONS AND NEXT STEPS

This final chapter identifies policy implications and next steps that policymakers must consider when implementing the assessments.

ASSESSMENTS COULD BE ROLLED OUT ON A TRIAL BASIS

The results of the field testing highlight that the science and art of measuring public health emergency response capabilities remain in their infancy. The assessments were found to be useful and feasible, but it is clear that data collected across different sites might not be fully comparable. Yet, even imperfect assessments of capability provide an important complement to the current practice of measuring preparedness according to plans, staffing, and equipment. Thus, it seems reasonable to consider rolling out the assessments on a trial basis. Given the issues raised in the field testing, however, CDC should put into place systems for collecting, learning from, and applying lessons from initial rollout.

Initial Rollout Should Include Efforts to Capture And Apply Lessons

CDC should consider the following steps to capture and apply lessons from initial rollout of the assessments.

Collect Data on Performance Metrics During the First Year of Implementation

To continue the process (begun with our initial field tests) of refining the measures, CDC/DSNS should collect and analyze data from the first-year rollout. This will require that awardees not only collect but also report data on the metrics. To facilitate this process, CDC/DSNS must be clear about whether state-level awardees should report data aggregated to the state level, data at the local, health-department level, or some combination. Specifying clear data-archiving practices up front will help ensure consistent and comparable data across time and location.

Additionally, CDC/DSNS should consider developing a standardized and concise feedback form to capture input from awardees about the clarity and feasibility of the tools. CDC/DSNS should consider collecting this feedback via the same portals as the performance-metric data, which should limit the additional burden on awardees. As noted above, the burden may be further limited by coordinating with CDC’s Division of State and Local Readiness (DSLR) data-collection and measure-development efforts.

Finally, as noted earlier in the report, minimizing assessment burden was a key driver in the development of the assessments described herein. Moreover, future decisions about the selection, development, and implementation of future SNS (and all PHEP) measures
should seek to determine whether the informational and evaluative benefits of collecting performance data justify the opportunity costs associated with the assessment. Thus, along with collecting performance information, CDC should consider collecting information from jurisdictions on the costs associated with using these assessments.

**Analyze Data Collected During the First Year and Revise Assessments Accordingly**

The most critical question about the measures is, Do they, indeed, get at critical drivers of preparedness and response? Thus, the ultimate goal of data analysis should be to assess predictive validity by comparing metric data, if possible, with results from actual incidents and events. Doing so, however, will require developing a specific methodology for comparing exercise-based and event-/incident-driven performance. Thus, the remainder of this section outlines analyses that can be accomplished more readily with existing methods and tools.

CDC/DSNS should analyze the data collected during the first year to assess the data quality and extract lessons learned. Data distributions should be examined for unexpected variance or outliers, and limited follow-up should be undertaken on cases with implausible values. Diagnosis of sources of measurement error (which could be done via telephone interviews) would improve data quality and provide insight into the data-collection process. The overarching goal should be to distinguish true score variance from error variance (especially systematic error).

Specific questions about the robustness of the metrics to variations in health-department structure and drill processes can be addressed by examining the extent to which metric values depend on these factors. Again, consideration should be given to separating differences in underlying capabilities from other sources of variation.

Once data quality has been considered, CDC/DSNS should seek opportunities to use the assessment metrics to test whether anticipated relationships among variables are, in fact, observed (a component of construct validity). This testing includes whether multiple measures of the same process agree with each other, and whether related constructs correlate sensibly (e.g., values on call-down drill metrics vary with the extent of prior notice).

**Develop Strategies for Ensuring that Awardees Use the Metrics to Improve Performance**

Ultimately, the purpose of performance measures is both to reveal performance gaps and to support efforts to close those gaps. Thus, efforts to refine and improve these measures should continue to focus on their potential to support quality-improvement (QI) efforts (Lotstein et al., 2007; Seid et al., 2007).
Ideally, the measures would be validated before they are used for QI. Given that the field of public health preparedness is in its early stages and that policy discussions often outpace science, we recommend that CDC/DSNS encourage initial use of the measures for QI, even while additional validity assessments continue. First, doing so will help ensure that the measures are seen right from the beginning as tools that can help jurisdictions improve, not just as a source of embarrassment when gaps are revealed. Second, QI efforts can provide useful sources of information about the quality of the measures themselves.

Specifically, we recommend that CDC/DSNS pursue a more focused pilot study of the measures, targeting a small number of awardees and involving the collection and analysis of data over time. Such a pilot study will provide a stronger validation opportunity for the metrics themselves and will allow for the exploration of how performance data can help to improve systems.

**Continue to Develop the Evidence Base Behind SNS Measures and Processes**

The data collected for these metrics, as well as the structure and process variables, will provide a valuable tool for more in-depth and scientific explorations. For example, one research question that might be addressed is, Do capacity measures on the TAR predict actual performance on operational-capability metrics? If so, Which ones are most predictive? Findings from such an inquiry could be used to test the validity of current models of preparedness and provide findings on which actions can be taken for targeting and improving training. Such efforts could complement the work of the newly reauthorized Centers for Public Health Preparedness, which are now focused on research, and of other efforts that follow from development of the PAHPA-mandated HHS Secretary’s research agenda.

**FUTURE MEASURE DEVELOPMENT EFFORTS MUST ADDRESS KEY ISSUES**

Beyond these immediate next steps, CDC and other decisionmakers will need to address a number of issues that pertain more broadly to capability measures for public health emergency preparedness.

**Balancing Standardization and Flexibility**

Observations from the field tests of the countermeasure-delivery capability assessments suggest that the assessments, with the possible exception of the computer-based mass dispensing drill, are quite feasible for most jurisdictions. However, the findings also suggest that further efforts will be required to control sources of variation captured by the assessments that go beyond the operational capabilities they were designed to measure (also known as *error variance*). For example, our observations suggest that measured performance might have been influenced by situational factors, such as differences in exercise scenarios, the time at which the assessment is conducted (i.e., whether exercises
are conducted during regular business hours), and the quality of the scripts used for staff-call-down drills.

Two possible approaches to addressing these issues are (a) to prescribe more specifically the drill scenarios and tasks, and (b) to collect information on drill context and processes and use this information to make post-hoc adjustments in interpreting assessment results. The assessments we created have purposely avoided prescribing specific exercise scenarios or procedures. As noted in the earlier RAND report (Nelson et al., 2007), a key design criterion that came out of discussions with CDC, state, and local SNS staff was to allow jurisdictions as much flexibility as possible. Thus, wherever possible, the assessments have avoided prescribing specific exercise scenarios or procedures. Instead, these assessments favor collecting extensive information on exercise scenarios and procedures to allow analysts to construct “apples-to-apples” comparisons and/or to statistically adjust assessment results to facilitate interpretation. Excel-based data-collection tools have been provided to facilitate collection of contextual information and calculation of key metrics.

However, the success of any post-hoc adjustment strategies will depend on the number of jurisdictions reporting data. Generally, fewer data points will limit analysts’ ability to distinguish signal from noise. Success of the post-hoc strategy will also depend on the ability to identify important contextual variables and on the amount of natural “spread” (variation) in the data (which will affect analysts’ ability to detect the signal in the data).

As a result, CDC/DSNS should be aware of the risk that even the fairly extensive amount of contextual information in the Excel-based tools might be insufficient to account for situational factors. Therefore, the agency should also consider whether greater standardization in scenarios and procedures is possible. This decision, of course, is not merely a technical matter but should also include judgments about the extent to which CDC is willing to impose what might be viewed as more-burdensome and more-prescriptive requirements on its awardees.

**Technical Assistance Requirements of Performance Assessments**

Another way to improve measurement quality is through technical assistance and training for those using the assessments. The assessments are designed to be used by jurisdictions without external assistance. Most of the assessments appear to be quite feasible; however, it is likely that training will be required to ensure consistent application of the procedures and metrics provided in the manuals.

The computer-based dispensing assessment will likely require substantially more technical assistance than the others. Expertise with computer modeling varies among health departments. Policymakers should consider whether to invest in providing more training and technical support to state and local health departments that are not familiar with the use and utility of computer models. This training would need to include how to use the software, and also how to interpret the results and use models in conjunction with
exercises in a cycle of continuous improvement. In some cases, health departments may need to consider hiring a consultant to manage the more-technical aspects of the assessment process. For states that have many small jurisdictions, it may be more efficient to have a state-level or regional “assessor” rather than trying to build that capacity into each individual local health department.
ANALYSIS USED TO SELECT SNS ELEMENTS AROUND WHICH TO BUILD ASSESSMENTS

We used critical path analysis (see, for instance, McClain, Thomas, and Mazzola, 1992) to identify the rate-limiting factors in the SNS process -- those tasks and activities that, if not done in a timely fashion, could prevent delivery of medications to the public within the 48 hours required by the Cities Readiness Initiative guidance. Examination of these factors helped identify the specific capabilities that are represented in Chapter 2’s conceptual model (see Figure 2.1).

The results of the analysis are summarized in the precedence diagram in Figure A.1. Precedence diagrams show the temporal flow of key tasks in a process and help to illuminate the most critical process elements. Each box in the diagram represents an SNS-related activity, with some activities carried out in sequence and others in parallel. Arrows indicate which functions are predecessors of other functions. For instance, the initial SNS shipment must arrive before the RSS can unload and repackage the shipment, which must occur before the distribution process can deliver materiel to the PODs, which must occur before PODs can begin dispensing.20

The diagram is organized into horizontal streams of function-specific activities, including Receipt, Staging, and Storage (RSS); Distribution; Dispensing (POD); and Public Information. We have also included a stream of activities related to inventory management; however, we recognize that some of the activities will also involve command and control (e.g., selection of dispensing sites). We have also included prophylaxis of local staff from local caches. Although not part of the SNS program, prophylaxis is nonetheless an important element of an SNS response.

The precedence diagram is also useful in identifying key SNS capabilities that are candidates for drill development. Crosscutting capabilities are more-general response elements that can be assembled and combined in order to execute a variety of tasks. For example, setting up a POD requires a solid command and control function, the ability to call down and mobilize staff, and so on. Visually, crosscutting capabilities are those activities that appear in multiple activity streams. These are highlighted in Figure A.1 by

20 In an ideal SNS response, the total time from request to dispensing would be limited only by the physical time it takes to move the SNS materiel. However, delays in any of the activities that parallel the tasks on the critical path could become rate-limiting factors and delay the response.
the vertical, solid-line,\textsuperscript{21} ovals (more detail on potential drill operations is provided in the earlier report).\textsuperscript{22}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figureA1.png}
\caption{Precedence Diagram with Crosscutting Capabilities Highlighted}
\end{figure}

In addition to these generic crosscutting capabilities, the precedence diagram also helps identify function-specific capabilities, highlighted in the diagram by the ovals. These capabilities are those activities that are “directly facilitate the dispensing of drugs and vaccines,”\textsuperscript{23} including requesting SNS assets, receiving of the SNS cache, distributing the materiel to Point of Dispensing sites or other dispensing nodes, actual dispensing of the materiel, and so on.

\begin{itemize}
\item \textsuperscript{21} The dashed oval around set up indicates that application of this capability will differ more substantially across RSS, Distribution, and PODs than other crosscutting capabilities.
\item \textsuperscript{22} Note that the precedence diagrams do not include the re-ordering process. Unfortunately, we have not had time to fully understand this process and have not included it as a candidate for drill development. For instance, it is not yet clear to us whether one should test the parts of the reordering process or whether it should be tested as a whole.
\item \textsuperscript{23} This phrase is taken from a ConOps statement from AHRQ on the design and operation of dispensing/vaccination clinics (Hupert et al., 2004). The document distinguishes “core” and “support” functions, which is roughly analogous to our distinction between “core SNS functions” and “crosscutting capabilities.” We believe that the term crosscutting is more apt and serves to highlight the extent to which different core functions draw upon a largely common pool of capabilities. Similar distinctions are also found in the literature on Six Sigma and other business-improvement packages (see, e.g., Pande, Neuman, and Cavanaugh, 2000).
\end{itemize}
Figure A.2: Precedence Diagram with Function-Specific Capabilities Highlighted

The results of this paper-and-pencil exercise were validated by discussions with subject-matter experts, practitioners in state and local health departments, examination of after-action reports, and direct observation of SNS drills and exercises.
APPENDIX B

Questions for Eliciting Feedback on Field Tests of SNS Capability Assessments

Following the field-testing of each assessment, RAND staff interviewed members of the health department using this semi-structured protocol to get feedback on the ease of use and the usefulness of the assessment tools. The questions were focused around four main areas: feasibility, reliability, validity, and utility.

Feasibility
1. How much additional effort/cost was involved in using the metrics operations manuals and collecting the metric data?
2. Were there any activities required by the metrics operations manuals that were particularly burdensome?
3. Are there ways that the burden could be lessened?
4. Was the burden reasonable?
5. Might the burden be different for other jurisdictions? If so, please explain.

Reliability
6. Where there any procedures or instructions in the metrics operations manuals that you found confusing?
7. Might another group of players/Controllers/Evaluators (in this or other jurisdictions) interpret any of the instructions differently?
8. Would it be possible to have another group of staff in the department (who have not been involved in discussions with us) review the manual to provide a fresh perspective on the manual’s clarity?

Validity
8. Do the metrics comport with other readiness assessments you’ve done? Can we examine documents (e.g., after action reports, or AARs)?
9. How does what you learned from this metric relate to what you’ve learned in previous TAR reviews?
10. Does the metric provide a fair and accurate picture of your capabilities?
11. How might the metric be revised to provide a clearer or more accurate picture?
12. Some have suggested that drills/exercises should be used for a “no-blame, no-shame” examination of problems, and that including them as part of a summative/accountability assessment system undercuts that goal. What do you think of this concern?

**Utility**

13. Does the metric provide information that you could use to improve your processes and systems?

14. Can you see any way in which the assessment might discourage process improvement?
REFERENCES


