Measuring and Assessing Public Health Emergency Preparedness: A Methodological Primer

Michael A. Stoto, Christopher D. Nelson, and the LAMPS investigators[1]

Introduction

Motivated by concerns about terrorism and natural disease outbreaks such as SARS and pandemic influenza, the U.S. federal government has invested more than $21 billion in the last decade to help state and local health departments prepare for large-scale emergencies such as bioterrorism and pandemic influenza.[2] In addition, governmental health agencies at all levels, hospitals and health care providers, and many other organizations have reviewed public health laws; developed and tested new emergency plans and procedures; purchased supplies and equipment; hired additional staff; and trained existing personnel to assume new roles and responsibilities when significant threats to the public health occur. Together, these investments have been part of a major concerted national effort to enhance and improve "public health emergency preparedness" (PHEP).

Given the magnitude of these investments, elected officials and policymakers need to know whether these efforts have actually substantially improved the nation's ability to respond to large-scale public health emergencies. In addition, if the nation's public health system remains underprepared (which is likely for some preparedness functions), policymakers and public health practitioners need guidance on how best to further improve the PHEP system's capabilities.

There has been considerable progress in measuring PHEP during the past decade. Increasingly transparent and systematic methods for measure development have been

[1] LAMPS (Linking Assessment and Measurement to Performance in PHEP Systems) is the CDC-funded Preparedness and Emergency Response Research Center (PERRC) based at Harvard School of Public Health. The LAMPS researchers who contributed to this paper through discussions and comments on earlier drafts include Paul Biddinger, Stan Finkelstein, Donald Goldmann, Melissa Higdon, Tamar Klaiman, John Kraemer, Richard Larson, Tara McCarthy, Rachael Piltch-Loeb, Elena Savoia, Ying Zhang, and K. Viswanath (PI). This paper has benefited from the contributions and comments on previous versions from many colleagues at Harvard and RAND, researchers associated with other PERRCs, and others. This white paper was developed with funding support awarded to the Harvard School of Public Health under cooperative agreements with the US Centers for Disease Control and Prevention (CDC) grant number 5P01TP000307-01.

[2] Source: Authors' calculations based on data for FY 2002-FY 2012 in the 2011 Trust for America's Health report (TFAH, 2011). This includes the total budget of the CDC Office of Public Health Preparedness and Response (and its predecessors), which primarily supports PHEP Cooperative Agreements with the states and selected local areas, and the SNS program, as well as the Hospital Preparedness Program now in the DHHS Office of the Assistant Secretary for Preparedness and Response (ASPR). It does not include more than $2 billion in state and local pandemic grants in FY 2006 and FY 2009 or more than $12 billion expended for development and federal purchase of influenza vaccines or other medical countermeasures. State and local funds and expenditures of other federal agencies are also not included.

developed for some of the major federal PHEP grant programs, including the CDC's PHEP Cooperative Agreement and the Strategic National Stockpile (SNS) program. During the early years of these programs, there were typically many measures included in the guidance, but little resulting data. Today, both programs collect, report, and use data from a focused set of measures developed using fairly transparent, replicable processes.

In spite of this progress, there is no clear scholarly and professional consensus on what good PHEP measures should look like or what methods are most likely to generate them. Previous assessments have shown little consistency in what constitutes "preparedness" or how it should be measured, and most instruments rely primarily on subjective capacity measures, lack scientific evidence, and fail to clearly define what entity is accountable for accomplishing the task or function (Asch, 2005; Nelson, Lurie, Wasserman 2007b). Indeed, the Institute of Medicine's (IOM) 2008 concluded that "it is difficult to measure objectively the progress that has been made," and four years later, "the preparedness gaps" remain.

Because of this difficulty, the IOM concluded that the future of public health preparedness requires "validated criteria and metrics that enable public health systems to achieve continuous improvement and to demonstrate the value of society's investment" (IOM, 2008). The IOM's call to action requires new quantitative and qualitative approaches to measuring public health systems' activities and associated outcomes, and to assessing whether health systems' performance meets the relevant standards.

Assessing public health system preparedness is challenging for many reasons. First and foremost, serious public health emergencies are (thankfully) rare, and play out differently depending on the context in which they occur. As a result, there are *few opportunities to assess outcomes by direct observation*. Statistical performance measures (such as the number of myocardial infarction patients who receive aspirin when they arrive at the emergency department) that hospitals regularly report on a weekly or monthly basis for accountability and quality improvement (QI) are simply not available for PHEP. Moreover, the uniqueness and relative infrequency of public health emergencies also make it difficult to learn from experience about what activities are more effective in improving preparedness.

Second, the range of activities that constitute an effective public health emergency response is complex and multifaceted, so even after the fact it can be *difficult to know what would have been the best response* to a specific situation given the absence of the "counterfactual" (what would have happened had some other response been undertaken). An effective approach in one community may be less so in others. For instance, during the response to 2009 H1N1, local jurisdictions were faced with decisions about closing schools: whether to do so at all, and if so, when should schools close and reopen. Their responses, decision processes, and even their stated rationales varied widely (Klaiman et al., 2010), and there is still no widely accepted understanding of whether school closures affected the spread of the virus or protected children. As a result, it is not clear either what aspects of preparedness should be measured to ensure

better results in the future or what levels of performance are required. For instance, would greater laboratory capacity have enhanced the response to 2009 H1N1? If so, how much would throughput or timeliness need to have been improved?

Third, public health systems themselves are multi-jurisdictional, multidisciplinary, and *system-level preparedness can be more or less than the sum of the parts*. The public health infrastructure (which itself is very heterogeneous across the nation) includes city, county, regional, and state health departments and offices as well as federal agencies. In addition to the basic infrastructure, "systems" that support public health preparedness include partner agencies such as hospitals and physicians; emergency medical services agencies; agricultural and environmental protection agencies; police; and others who may not think of themselves as having a public health role. As a result of all of these factors, responsibility and accountability for public health preparedness is diffuse, making it difficult to determine which partner's performance to measure, and how to hold each partnering entity accountable for its contributions.

In response to the IOM's call for the development of better methods of measurement, the CDC awarded a grant to the Harvard School of Public Health's Preparedness and Emergency Response Research Centers program and its partners focusing on PHEP measurement issues. With this funding, LAMPS (Linking Assessment and Measurement to Performance in PHEP Systems) investigators at Harvard and partnering institutions are working to develop valid and reliable criteria and metrics to assess and ultimately improve public health emergency preparedness in the United States.

The ultimate goal of LAMPS is to develop valid, reliable, and practical PHEP measures that can be used to ensure accountability for a decade's worth of PHEP investments, to improve public health systems' capabilities to respond effectively to future public health emergencies, and to support related research efforts. To this end, this paper will summarize insights on PHEP measure development that are emerging from this effort, drawing on the experience of the authors and their public health practice partners as well as on discussions in workshops, exercises, and other venues, only some of which has been published in scientific and professional journals. The goal is <u>not</u> to describe a universal set of preparedness measures, but rather an approach that others can use to develop measures as needed for a variety of purposes.

Adopting a public health systems research perspective, we begin by describing an approach to the development of performance measures and assessment tools that has proven to be effective in health services research and healthcare quality improvement efforts. We then apply this approach to PHEP, illustrating each of the four major steps in the measurement development cycle with preparedness examples, many of which are drawn from the work of the LAMPS team. A companion paper will use this framework to assess the current state of PHEP measurement in the United States, identifying areas in which more measurement development work is needed.

A Public Health Systems Framework for Measuring and Assessing PHEP

Realizing the full potential of PHEP improvement initiatives first requires a coherent, robust, and integrated national performance measurement system. In a report on healthcare, *Performance Measurement*, for instance, the IOM has noted that a measurement system, regardless of the area measured, should link performance measures directly to explicit national improvement goals in a manner that is purposeful, comprehensive, efficient, and transparent. Specifically, the IOM notes that a national system should possess the following attributes: specific purposes and aims; a plan for the development and promulgation of performance measures; a system of data collection, data validation and aggregation processes; a system of public performance reporting in support of decisionmaking activities; funding for research activities; and a continuous evaluation process and impact assessment of performance measurement and quality improvement initiatives. Measures should be comprehensive, applicable to longitudinal assessments, based on a system-level approach, and designed to address shared accountability issues regarding system improvement and research. Such performance measures should provide a benchmark for an acceptable minimum set of standards and an agenda for improvement for each community, while allowing them the flexibility to establish additional, locally relevant goals (IOM, 2007). Performance measures can be quantitative, qualitative, or a combination of the two.

Drawing on an extensive literature in health services research, healthcare quality improvement, and other fields (Association of Public Health Observatories, 2008; National Quality Forum, no date) the IOM performance measurement report provides a solid platform for PHEP measurement development efforts. For instance, applying Donabedian's (2005) "structure-process-outcome" model to the PHEP system would help to establish the link between structures (capacities such as resources and trained staff), processes (capabilities such as preparedness communication or the ability to implement social distancing measures), and outcomes (effects such as mitigating mortality and various negative health, psychological, and social consequences of health emergencies).

This "science of measurement" is incorporated in the National Quality Forum's (NQF) measurement evaluation criteria. NQF, a nonprofit organization created in 1999 to improve the quality of American healthcare by building consensus on national priorities and goals for performance improvement and working in partnership to achieve them, endorses national consensus standards for measuring and publicly reporting on healthcare system performance using the following four criteria:

- *Importance to measure and report:* The extent to which the specific measure focus is evidence-based (for instance, evidence that performing the process measured will help achieve the intended outcome), and important to making significant gains in healthcare quality and improving health outcomes for a specific high-impact aspect of healthcare where there is variation in or overall less-than-optimal performance.
- *Scientific acceptability of the measurement properties:* The extent to which the

measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.
- *Usability and use:* The extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) can understand the results of the measure and find them useful for both accountability and informing quality improvement.
- *Feasibility:* The extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement fields (National Quality Forum, no date b).

The health services literature indicates that an effective set of performance measures is best developed through an iterative measurement development cycle that addresses four key questions: (1) Why measure (clarification of the purposes and uses of the measurement effort); (2) What to measure (identification of the concepts, domains, or specific activities and processes to be measured); (3) How to measure (development of specific indicators or metrics and measurement systems); and (4) How well the metrics work (assessment of validity, reliability, practicality, and utility).

In addition, PHEP assessments should capture the "systems" nature of the preparedness enterprise, which requires not just measuring the most critical system components, but also representing how they interact with each other and with their contexts to produce outcomes (see, for example, Simon, 1981). Following the definition of key terms, we use this framework to organize this section, proceeding through each of the steps with examples from LAMPS research and other efforts.

Defining key terms

Developed in an *ad hoc* fashion, current terminology used to describe PHEP measurement efforts is often inconsistent, so it is important to begin with definitions. And while the IOM report described above focuses on "criteria and metrics," we find it useful to use the broader terms *measurement* and *assessment*, which we understand to include a variety of components summarized in Figure 1 (adapted from Nelson, Lurie, & Wasserman, 2007).

Metrics or measures are the focal points of measurement and assessment systems that define the specific aspects of PHEP performance (e.g., observable behaviors, processes, and attitudes) of interest and to varying degrees prescribe the ways in which these aspects are described (ranging from definitions of key terms to formal scales and rating systems). In particular, we distinguish below between measures of preparedness *capacities* and response *capabilities*. At the most basic level, metrics seek to provide a common "lingua franca" that can be used to describe and evaluate performance. Metrics appear near the bottom of the conceptual framework in Figure 1, however, because they are actually the product of many other elements.

For instance, to be meaningful, metrics must be populated with information generated by some data-generating *measurement process*. Data results from observations of performed tasks (written or operational), as described through surveys or checklists,

focusing on capacities (e.g., plans, personnel, equipment) or operational capabilities (the ability to execute plans). Performance might be observed in drills, exercises, or actual public health emergencies, and be judged by participants or external observers. The result of this measurement process is a set of PHEP measures (which tend to be quantitative in nature) or assessments (which tend to be qualitative).

Measurement and assessment systems include a number of other elements that influence how performance data are created, interpreted, and used. For instance, measurement systems start from an understanding (which may be implicit) about the critical elements of preparedness and what constitutes quality in each domain. Performance *standards* are then created to help users determine which values on the metrics should count as "good" or "acceptable." In short, metrics are the observable "yardsticks" used to judge performance and standards are the thresholds that define how "good" is "good enough" on those metrics (Nelson, 2010a).

The selection of preparedness elements and performance standards is informed by an understanding of the *PHEP evidence base.* In clinical medicine, for instance, the performance metrics selected (e.g., the proportion of myocardial infarction patients receiving aspirin on arrival in the emergency department) based on RCTs and other studies demonstrating a causal link between the activity measured and some desirable outcome (e.g., reduced morbidity and mortality). At the other extreme, the evidence base might be the collective experience of a group of individuals developing metrics or interpretations of past responses.

All of this, of course, relates to the purposes or uses of the performance assessments, which range from accountability to system improvement to research. The purposes or uses of performance measures are in turn influenced by users' incentives and resources (see Stecher, 2010). Indeed, clarifying the reasons why performance measures are needed is the first stage of the measurement development process, which we now discuss.

Why measure: Clarifying the purposes and uses of the measurement effort

Three purposes drive the development of PHEP measures: accountability, systems improvement, and advancement of knowledge (research). The most prominent use in a given situation will influence the choice and specification of metrics. In their conceptual framework for assessment, Nelson and colleagues (2007a) stress the close link between assessment, purpose, and performance when they state that "preparedness involves a coordinated and continuous process of planning and implementation that relies on measuring performance and taking corrective action."

Decisions about the intended uses and users of assessment systems often have important implications for the construction of specific measures. As Nelson, Lurie, and Wasserman (2007b) note, performance improvement efforts are best supported by metrics that are highly relevant to issues of specific concern to a state or local health department (or other entity being evaluated). Accountability decisions and research

enterprises, by contrast, usually involve comparisons across jurisdictions, and thus require more standardization than improvement metrics.

Accountability. Policy makers need to know how well prepared health systems are, and whether their levels of preparedness have been improving. Such information is necessary in order to set priorities and hold health departments accountable for PHEP investments. Indeed, the *Pandemic and All Hazards Preparedness Act* (PAHPA) requires the U.S. Department of Health and Human Services to develop performance standards and metrics. Starting in Fiscal Year 2010, a portion of CDC funding to state and local health departments has been tied to their ability to demonstrate that they have achieved a minimum level of preparedness based on such standards and metrics. CDC's performance measures initially focused on infrastructure development and completion of (but not performance of) operational assessments such as call-down drills. In future years it is expected that funding will be tied to *levels* of operational performance, defined as how well the unit performed on the drill.

Quality improvement. The U.S. National Health Security Strategy (NHSS) identifies quality improvement as one of ten strategic objectives and calls for more attention to systematic quality improvement methods and a national culture of quality improvement to enhance national health security. The "QI approach" typically employs a rapid Plan-Do-Study-Act (PDSA) strategy in which the "study" phase requires quantitative performance measures or qualitative assessments of what was accomplished in the "do" phase, which are then translated into further system improvements in the "act" phase. Lotstein and colleagues, for instance, have recently documented the importance of metrics for quality improvement efforts to improve PHEP, specifically relating to pandemic influenza (Lotstein, 2008).

Research. The science of public health preparedness is in its infancy, and much research is needed to fill in gaps in determining "what works." To be able to test the effectiveness of specific PHEP interventions, valid and reliable measures of PHEP outcomes, intermediate outputs, and inputs are needed for research studies. Since structure and process measures should focus on activities that are known to improve outcomes (as discussed in more detail below), the current paucity of high-quality research studies in PHEP (Nelson, 2007a; Nelson, 2008; Savoia, 2009b; Acosta, 2009) provides a challenge to PHEP measurement. Given aggressive measure-development timelines in the PAHPA legislation, Nelson and colleagues (2009) suggest that in the short-term we must be willing to employ a fuller range of evidence sources in developing "evidence-based" measures. These initial (and often imperfect) measures can then serve as a basis for improving the evidence base, ultimately leading to refinement of the measures over time.

What to measure: Identification of the concepts or domains to be measured

Measurement systems must begin with clear definitions of key concepts and result in identification of specific activities, processes, and attitudes to measure. Beginning at the most general level, we have adopted the definition of PHEP developed by Nelson

and colleagues through a consensus process, namely, "the capability of the public health and healthcare systems, communities, and individuals, to prevent, protect against, quickly respond to, and recover from health emergencies, particularly those whose scale, timing, or unpredictability threatens to overwhelm routine capabilities. Preparedness involves a coordinated and continuous process of planning and implementation that relies on measuring performance and taking corrective action" (Nelson, 2007a).

At the heart of this definition is the concept of "public health system" as a complex network of individuals and organizations that have the potential to play critical roles in creating the conditions for health. As developed by the Institute of Medicine, Figure 2 represents seven key sectors that can work individually or together as part of a public health system to create the conditions necessary for public health emergency preparedness, response, and recovery. While each of these actors is a separate entity, a robust public health system for preparedness requires that each work together when appropriate (IOM, 2003, 2008).

Implicit in the Nelson (2007a) definition is a fundamental distinction between PHEP capacities (akin to "structure" measures in the Donabedian framework), capabilities (comparable to "process" measures), and preparedness outcomes. *Capacities* represent the resources—infrastructure, response mechanisms, knowledgeable and trained personnel—that a public health system has to draw upon. Capacity-building activities such as planning, training, and acquiring equipment and supplies belong in the preparedness domain of what systems do to get ready for the next emergency. While minimum capacity levels are necessary elements, capacity alone is not sufficient to ensure preparedness.

Capabilities, on the other hand, describe the functional or operational actions a public health system is capable of taking to effectively identify, characterize, and respond to emergencies: surveillance, epidemiologic investigations, laboratory, disease prevention and mitigation, surge capacity for healthcare services, risk communication to the public, and coordination of system responses through an effective incident management. Capabilities, therefore, are latent characteristics of the PHEP system, part of the response rather than the preparedness domain. As such, capabilities are best measured and assessed in actual emergencies or other situations in which the PHEP system responds to an emergency.

Drills and exercises can reflect both capacities and capabilities. Viewed both as training and as opportunities to evaluate plans, they are capacity-building activities. Drills and exercises also make it possible to observe components of the PHEP system in action and learn about its response capabilities (Biddinger, 2008; Savoia 2009a).

Distinguishing between capacities and capabilities is one way to recognize the heterogeneity in the organization of U.S. public health systems. Depending on the division of responsibilities between state and local health departments, for instance, states employ different combinations of *capacities* such as disease reporting requirements and processes; syndromic surveillance systems that draw data

automatically from hospitals; a network of state, local and private-sector laboratories; and trained epidemiologists. Regardless of the approach they take, however, PHEP system preparedness requires that states have the capability to detect and characterize disease outbreaks in a reasonable time frame. This surveillance *capability* can be assessed in drills and exercises as well as in real events such as routine foodborne disease outbreaks or more serious events such as the 2009 H1N1 pandemic (Lurie, 2004; Lotstein, 2008; Zhang, 2011; Stoto 2012a). Because expectations regarding capabilities are more uniform than for the capacities needed to achieve them (which necessarily reflect differences in state and local public health systems), capability measures can be more appropriate for comparing states in an accountability framework.

In public health systems research, development of a logic model that specifies the critical goals and objectives of public health preparedness, as well as how various functions, processes, and resources contribute to meeting them, is necessary prior to the development of specific measures. One such logic model (Figure 3, adapted from Stoto, 2005) specifies the goals and objectives of public health preparedness, and the capabilities and capacity-building activities intended to achieve those goals and objectives. This model can help clarify how various capabilities contribute to the overall goal of assessing and improving preparedness and how a variety of public and private community organizations can contribute to overall public health preparedness. Causal relationships or “drivers” generally go from left to right, but the actual relationships between capacities, capabilities, and outcomes depend on context, both the specific emergency and the structure of the PHEP systems responding to it.

In this model, the capabilities are described within the legal, economic, and operational (LEO) domains suggested by Potter and colleagues (2012). In addition, we have added “social capital” to describe the intangible partnership and informal relationships between individuals and organizations that are critical to effective emergency operations (Koh, 2008; Stoto, 2008).

The three main functional capabilities correspond to the three core functions of public health (IOM, 1988), and represent what the public health system must accomplish to respond effectively. The fourth capability–coordination and communication–represents a series of interrelated functions needed to ensure that the system fulfills its assessment, policy development, and assurance roles: communication with the public, information sharing within the public health system, and incident management and leadership.

Continuing the discussion of the distinction between capabilities and capacities above, note that doing drills, exercises, and after-action reports (AARs) is represented in the PHEP logic model as demonstrating operational capacities. Such activities can also be opportunities to measure and assess a PHEP system’s capabilities.

Before specific metrics can be developed, consideration must first be given to identifying the aspects of preparedness that need to be monitored. For instance, before an emergency occurs, it is critical to understand the consequences of alternative PHEP improvement strategies in order to determine which capabilities (the “domains” or major concepts included in a measurement system) and how many of these capabilities (the

"criteria" or thresholds) are necessary. Below we identify a number of sources of evidence and thought processes that can be used to identify specific points of measurement.

Scientific literature

The most traditional, evidence-based approach to identifying critical dimensions of PHEP measures is represented by Parker and colleagues' development of a model for crisis decisionmaking in public health emergencies (Parker, 2009). Rather than critiquing the decisions themselves, this model focuses on the decisionmaking process that connects decision inputs (such as people, systems, and organizational structure) with decision outputs (i.e. decisions) to take particular actions. Critical to all decision processes, the authors note, are situational awareness, action planning, and process control. Using these as an organizing structure, Parker and colleagues have developed a paper-and-pencil assessment form intended for use in exercises and real incidents to allow public health practitioners to assess their baseline crisis decisionmaking capabilities and identify shortfalls and shortcomings that may represent opportunities for internal process improvements.

For instance, in order to identify the most salient factors associated with H1N1 vaccine uptake in the United States, Galarce and colleagues began by conducting five focus groups with participants from diverse ethnic/racial and socioeconomic backgrounds. Viewed in the light of existing theoretical models of communication, the focus group results identified key themes that might affect vaccine acceptance, such as H1N1 knowledge, preventive behaviors, attitudes, beliefs, mass and interpersonal communication, and emergency preparedness in general (Galarce, 2011).

Process observation and mapping

However, scientific literature is often not available for many aspects of PHEP, or is insufficient to fully support selection of points of measurement. One approach that can be useful in such circumstances is process observation and mapping. Nelson and colleagues (2009) illustrate this approach in their development of performance measures for the Strategic National Stockpile (SNS). The precedence diagram reproduced in Figure 4 provides a high-level representation of the SNS delivery process, beginning with the request of SNS materiel and proceeding from distribution from central warehouses to Points of Dispensing, where the materiel is provided to individuals. The diagram helps distinguish sequential from parallel processes.

Based on observations of SNS exercises and in-depth interviews with federal, state, and local practitioners, Nelson and colleagues identified a number of cross-cutting and function-specific capabilities on which they focused their measure development activities. *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 command and control function, the ability to call up and mobilize staff, etc. Visually, *crosscutting* capabilities are those activities that appear in multiple activity streams (see the vertical, solid-line, ovals). *Function-specific capabilities* are highlighted in the diagram by the horizontal ovals. These capabilities are those activities that

"directly facilitate the dispensing of drugs and vaccines," including requesting SNS assets, receipt of the SNS cache, distributing the materiel to Point of Dispensing sites (PODs) or other dispensing nodes, actual dispensing of the materiel, and so on.

Computer simulations

In some instances, understanding of key response processes can be represented mathematically, allowing researchers to identify potential points of measures by observing which parameters in the model appear to offer the most leverage over outcomes. For instance, researchers at the University of Pittsburgh PERRC have used simulation models to study different approaches to school closing to prevent the spread of pandemic influenza in the community. They found that (1) statewide, centralized school closure authority is not necessarily optimal for reducing influenza attack rates, and that school-by-school decisions based on the school's own cases can be better; (2) school closures must be long and consistently sustained to reduce influenza attack rates; and (3) school closures implemented too early in a pandemic are not effective in reducing influenza attack rates (Lee, 2010). These results suggest the need for indicators of a state's school system's ability to implement a school closing policy that allows for school-by-school decisions, is not triggered too early, and can be maintained long enough to be effective. This, in turn, suggests that public health systems (including schools) need to measure their surveillance capabilities at individual schools to identify cases and their ability to engage the relevant parties to make school-by-school decisions about closing and re-opening.

Eliciting practitioners' insights

A final approach that has been recently used with success involves eliciting practical insights from individuals with experience preparing for and responding to public health emergencies. For instance, Nelson and colleagues used a formal expert panel process to develop recommended POD infrastructure standards (Nelson, 2010a). Similarly, the CDC has convened a PHEP Evaluation Workgroup (PHEP-EWG) to guide the overall process and provide "big-picture" guidance on the full range of PHEP capabilities and capability-specific subgroups to provide subject matter expertise. The PHEP-EWG consists of a core group of federal, state, local, and non-governmental partners and stakeholders with expertise and interest in PHEP measurement and evaluation. Subgroups consisted of external partners with one or more of the following: measurement expertise, content expertise, and in-the-field experience. To ensure consistency across the Workgroup and subgroups, several Workgroup members also served on the subgroups. These groups provided guidance in the use of all other methods, in some instances helping to create process maps, synthesize existing evidence, or share their own practical experiences (CDC 2009; Shelton, 2011).

Practitioner elicitation can be combined with any and all of the other approaches. For instance, the panel convened for the POD standards was informed by reviews of scientific literature, surveys of current practice in large cities, and results of mathematical modeling of POD operations and POD location decisions (Nelson, 2010a). Similarly, discussions by the CDC's PHEP-EWG often featured collaborative

efforts to develop process maps (CDC 2009; Shelton, 2011).

How to measure: Development of specific indicators or metrics and measurement systems

The health services research/public health systems research perspective distinguishes between general concepts and specific measures (or metrics) of those concepts. Concepts such as the ability of public health systems to detect outbreaks and respond effectively, or to communicate with vulnerable populations, can be operationalized in many ways. As noted by the IOM (2007), a practical and useful set of performance measures requires a system of data collection, data validation, and an aggregation of processes. As described in Figure 1, having identified uses for the metrics and the key concepts to be included as discussed above, the next step in developing PHEP measures requires developing valid and reliable measures of the major concepts, identifying sets of measures that together reflect the key dimensions of preparedness, and developing a measurement process to operationalize the metrics.

Developing measures of specific preparedness concepts

For accountability purposes, PHEP measures must be standardized so that they are comparable over time and between units, or against some desired performance threshold, as described in Figure 1. Definitions include specification of the unit of measurement (individual, as in a measure of an emergency response worker; team, as in the group of individuals who staff a countermeasures distribution site; state or local health department; entire community; and so on), time frame, and similar matters. Defining the unit of measurement can be difficult, especially since some PHEP functions are normally executed at the local level and others at the regional or state levels (Koh, 2008; Stoto, 2008). Approaches to “rolling up” measures from local to state health departments, for instance, should also be defined in a standardized way. If the measure tracks the amount of time required to accomplish a critical task (“time to” measures), what it means to accomplish the task, as well as start and stop times, must be standardized. For measures that are to be reported on a per capita basis, consistent population counts are needed for the denominator.

Many PHEP measurement systems elicit knowledge, attitudes, or practices of emergency workers, the general public, or others through survey questionnaires. For such measures, as well as observation guides for drills and exercises, a large body of research and experience has helped to identify a set of good practices that should be employed for PHEP measures. For instance, the San Francisco Bay Area Advanced Practice Center (2010) has compiled a Seasonal & Pandemic Influenza Vaccination Assessment Toolkit. These practices include careful specification of the concepts to be assessed, literature reviews to identify existing validated questionnaire items, pilot testing, cognitive testing, and statistical evaluation of the questionnaire’s validity and reliability (to be discussed further in the next section).

For instance, Galarce and colleagues (2011) built on the literature review and focus

groups described above to generate survey questions, which were combined with items adapted from existing instruments and further refined through cognitive interviews with potential survey respondents. Implemented in a nationally representative survey, they found that individuals' attitudes about H1N1 vaccination were strongly associated with age, urbanicity, perceiving the H1N1 vaccine as safe, and receipt of seasonal flu vaccination. These findings suggest that in future emergencies, public health communication campaigns be targeted towards audiences segmented by social class, race/ethnicity, and beliefs—what advertisers often call "psycho-demographics" (Galarce, 2011).

Similarly, Barnett, Balicer, and colleagues have used similar formal methods to develop and validate a series of survey questions to assess the determinants of individual local public health employees', emergency medical services workers', and hospital workers' willingness to respond to public health emergencies. Using Witte's Extended Process Model (Witte, 1992) as a starting point, they demonstrate the impact of perceived threat and efficacy on local public health workers' response willingness to public health emergencies (Barnett 2009, 2010; Balicer, 2010).

Standard procedures for the development of individual-level measures can also be adapted to assess system-level attributes based on individual observers. Biddinger, Savoia, and colleagues have used similar methods to develop and test measures of public health system preparedness based on participants' and external observers' assessment of performance observed in emergency preparedness exercises (Biddinger, 2008; Savoia, 2009a). This work is discussed further below in the section on assessing validity and reliability.

PHEP performance measures can be quantitative (e.g., the time required to complete a task, number of vaccine doses on hand) or qualitative (an expert assessment of the performance of a health department regarding surveillance procedures in a tabletop exercise). Qualitative assessment are often represented in quantitative forms, as in a 5-point scale (e.g., where 1 represents "response not sufficient" and 5 represents "response exceeded expectations").

Quantitative measures are seen as more objective, but their objectivity depends on who is doing the assessing, whether they have been adequately trained as evaluators, and other aspects of how the measurement system has been implemented. For instance, starting the clock later on "time to" measures can give the appearance of a shorter time needed to accomplish the task. Policymakers often take far too much comfort in numbers without asking about the process that generates those numbers. This is especially true since the PHEP concepts we are trying to quantify are often difficult to define precisely, so the potential for deviations from measurement protocols makes them less objective than they appear. Indeed, one must always be mindful of "Stamp's Law of Statistics," attributed to Harold Cox, which notes that the source of the data may easily be the weakest link. As a young man in India, Cox quoted some statistics to a Judge named Stamp. Judge Stamp responded: "Cox, when you are a bit older, you will not quote Indian statistics with that assurance. The Government is very keen on amassing statistics - they collect them, add them, raise them to the nth power, take the

cube root and prepare wonderful diagrams. But what you must never forget is that every one of these figures comes in the first place from the village watchman, who just puts down what he damn pleases" (Stamp, 1929).

Developing sets of performance measures

Beyond the qualities of the individual measures, consideration must be given to how a set or group of measures work together to assess PHEP. Andrews (1989) summarized the key characteristics of health indicators: "a limited yet comprehensive set of coherent and significant indicators which can be monitored over time, and which can be disaggregated to the level of the relevant social unit." Clearly this statement embodies tensions: between a limited number of measures that nevertheless comprehensively assess the critical PHEP concepts; between measures that are significant on their own yet together tell a coherent story; and between disaggregated measures (which may require small sample sizes) that nevertheless can be monitored over time (despite the variability due to small sample sizes) (Stoto, 1992; 2007). To guide healthcare quality improvement efforts, the Institute for Healthcare Improvement (IHI) recommends balancing these concerns by having a small number of measures that refer to agreed-upon goals and that the measures include a balance of outcome and process measures and together describe an exemplary system of care. IHI's "Whole System Measures," for instance, consist of 13 indicators developed to measure the overall quality of a hospital, group practice, or large healthcare system (Martin, 2007).

Review of the PHEP logic model can help in identifying a manageable number of measures that comprehensively cover the domains of interest. For instance, the Harvard School of Public Health Center for Public Health Preparedness (HSPH-CPHP) exercise program develops observational guides for evaluators by considering the major response capability domains in the logic model: surveillance and epidemiology, disease control and prevention, mass care, communication within the "public health system," communication with the public, and leadership and management. Then, depending on the scenario, particular aspects of each domain are chosen for measurement. For instance, in a pandemic flu scenario, the specific measures for the surveillance and epidemiology and disease control and prevention domains are illustrated in Box 2.

Many of the approaches discussed above in the "What to Measure" section can be used to determine the components of a set of measures. These include logic models, reviewing the scientific literature, process mapping, and computer simulations. The goal in such analyses is to identify the specific capacities and capabilities under control of the public health system that are most likely to lead to better outcomes in a variety of emergency scenarios. For instance, RAND and CDC adapted engineering concepts such as critical-path analysis 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 (Nelson, 2009).

One of the realities of performance-based management is that the organizations whose performances are being assessed tend to focus their attention on the aspects of their

performance that will be measured, perhaps to the detriment to other areas. To address this, the elements of PHEP measurement sets can be "rotated" over time, replacing measures on which most units have achieved success with others addressing aspects of preparedness in need of attention. Enough measures must be maintained, however, to allow for the measurement of improvement over time.

Operationalizing measurement systems

As suggested in Figure 1, once the individual and sets of measures have been identified, measures must be clearly operationalized with the required data elements explicitly detailed (Derose, 2002), and a consistent and reproducible measurement process must be set up to actually obtain the data.

The first step is to clarify the unit of measurement. For instance, a metric might refer to the capabilities of individual members of a health department's staff, the capacities of a state or local health department itself, or the performance of the entire PHEP system in a community or a state. Sometimes the unit of observation is different than the unit of measurement. Individuals may be surveyed about the agencies or the PHEP systems they work in, or agency leaders asked to report on the degree to which their staff is appropriately trained. Metrics defined at one level are sometimes "rolled up" to a higher one. Capacity surveys conducted at the county health department level, for instance, can be rolled up to the state level by averaging the county measures or by calculating the proportion of counties in a state that meet some threshold.

Many existing PHEP measurement systems ask whether health departments or other PHEP partners have undertaken various preparedness activities or met specified performance standards. For instance, the metrics required by CDC's PHEP cooperative agreement report on state and local public health laboratory capacities needed to respond rapidly, identify or rule out particular known biological agents, and increase the workforce and laboratory throughput needed to process large numbers of samples during an emergency (CDC, 2008). Some instruments assess whether state and local health departments have used resources provided by the federal government according to the program guidelines (Beitsch, 2006). Other capacity instruments are based on self-assessments of public health agencies (Lovelace, 2007), their staff members' competencies (Brand, 2006), or whether the jurisdiction and its staff have tested their plans through exercises and/or used the capabilities in question during real responses (CDC, 2008).

These assessments are "standards-based" in the sense of content standards: They prescribe the topics jurisdictions need to be proficient in. They vary in their use of performance standards–thresholds that define "acceptable" performance. Such surveys, inventories, and structured assessments often rely on self-reports from those being assessed, such as the annual reports filed by health departments that receive funding under the CDC PHEP cooperative agreement.

Performance might be observed in drills, exercises, or actual public health emergencies, with observations recorded through surveys or checklists. If the measure is to be obtained by observation of an exercise or actual event, the activities that "count" for

generating data must be clearly specified. If measurement is to be based on participant or external observers, exercise evaluation guides (EEG) must be developed as well as a process for the observers to complete them and for the resulting data to be gathered and analyzed. For EEGs and survey questionnaires, data might be gathered on paper forms and entered into a database for analysis afterwards, or respondents might be asked to complete a web-based questionnaire (which simplifies data entry and cleaning, but may not be practical in an exercise in which the observers don't have immediate access to a computer).

One of the open questions in PHEP measurement is whether to use external observers in exercises and actual events, as opposed to having participants rate their own experience. External observers have more objectivity and are presumably less likely to be biased in favor of the unit being observed. On the other hand, external observers may not know enough about the PHEP issues or the system they are observing to correctly interpret what they see, and engaging them can be costly. At a minimum, standards are needed to ensure that observers have the necessary training and experience. Perhaps a system can be developed in which PHEP planners are used as observers in other similar jurisdictions.

Qualitative assessments of system performance

When the focus is on improvement rather than accountability, and on complex PHEP systems rather than their components or individuals, qualitative assessment of the system capabilities of PHEP systems can be more useful than quantitative metrics. Ensuring that such assessments are rigorous can be challenging, but a well-established body of social science methods provides a useful approach. For example, based on discussions at an international symposium on Health Policy and Systems Research (HPSR), Gilson and colleagues (2011) summarize a series of concrete processes for ensuring rigor in case study and qualitative data collection and analysis (see Box 1). Because the focus is on systems rather than individuals, Yin's (2009) classic book on case study methods, now in its 4thedition, is also relevant. March and colleagues (1991) and Weick and Sutcliffe (2001) offer more specific suggestions relevant to PHEP. The realist evaluation perspective (Pawson and Tilley, 1997) also has much to offer. The following section summarizes some of the key ideas from this literature.

Use of theory. Gilson and colleagues (2011) recommend the use of theory to guide sample selection, data collection and analysis, and to draw into interpretive analysis. "Theory" in this context is broad, ranging from basic social science theories about risk communication to preparedness doctrine as embodied in the National Incident Management Strategy (NIMS) or the National Health Security Strategy (NHSS) to the context (C), mechanism (M), and outcome (O) configurations of realist evaluation (Pawson and Tilley, 1997) that attempt to specify "what works for whom in what circumstances and in what respects." For example, Stoto and colleagues (RAND 2005) developed a PHEP logic model (Figure 3) to focus their case studies of the public health response to West Nile Virus, SARS, monkeypox, and hepatitis A on core PHEP

capabilities (assessment, policy development, assurance, public communication, emergency management) in data gathering and analysis.

Case selection. Gilson and colleagues recommend a purposive–rather than random–approach to selecting cases to allow prior theory and initial assumptions to be tested or to examine both ‘‘average’’ and unusual experience. For instance, Klaiman and colleagues (2012) used a positive deviance approach to identify high-performing local health departments for our vaccination clinic study.

It can also be useful to conduct a series of parallel case studies, where cases are chosen intentionally to reflect a variety of settings such as health department types, geographical areas, and populations served; and to test theories about the determinants of effective PHEP systems (Yin, 2009, especially figure 2.2 on p. 39). Weick and Sutcliffe (2001) stress the importance of learning from “near misses,” which can reveal potential system problems that can compound to create serious problems. In a study of the public health response to West Nile Virus, SARS, monkeypox, and hepatitis A, which can be regarded as near misses since they were not catastrophic, Stoto and colleagues (RAND 2005) studied a sample of state and local health departments from a variety of settings and the degrees to which they experienced these public health emergencies.

Gilson and colleagues (2011) also recommend negative case analysis, i.e., specifically looking for evidence that contradicts current explanations and theory, and refining them in response to this evidence. For example, Stoto’s analysis of influenza surveillance systems offered a number of surprises relative to pandemic planning assumptions. In terms of the epidemiology, it was assumed that a pandemic viral strain would emerge in Asia and would be virulent in addition to easy to transmit; it was not assumed that children would be a higher risk for infection or severe consequences (Stoto 2012). As a result, school closing policies based on the premise that children were efficient vectors for community transmission lost their rationale when it appeared that children were the ones who needed protection (Klaiman et al., 2011).

Experience more interpretations. Particularly when dealing with single, unique cases, March and colleagues (1991) stress the need for multiple observers in order to increase the number of interpretations, which create a mosaic of conflicting lessons. In particular, Weick and Sutcliffe recommend that researchers resist the temptation to “normalize” unexpected events. In the face of system failure, they argue, it is natural to look for evidence that confirms expectations, which postpones the realization that something unexpected is developing. Rather, experience that calls into question planning assumptions is more important than experience that confirms them (Weick and Sutcliffe, 2001).The “facilitated look-backs”(Aledort, 2006) used in preparing the Massachusetts Department of Public Health 2009 H1N1 After Action Report (AAR) provided an opportunity to review events that occurred during the pandemic from the point of view of state and local health departments, healthcare providers, school systems, and others in order to explore different interpretations of events and perceptions of what happened (MDPH, 2011).

Experience more preferences. To learn as much from single, unique cases, March and colleagues (1991) recommend gathering as much information as possible on the preferences and values organizations use to distinguish successes from failures. Weick and Sutcliffe (2001) illustrate this point in their observation that nuclear power plants' departments can be observed in meetings questioning the interpretation of other departments, adding their own perspective on what's at risk. This interaction generates hypotheses about what is going on, what can be done, and what the long-term consequences of proposed actions might be. The diversity of expertise involved in this interaction decreases simplification and increases mindfulness by enabling people to see different things when the view the "same" event (Weick and Sutcliffe, 2001). For instance, drawing on the 2009 H1N1 experience, Nelson and Plough (2012) use a "grounded theory" case study approach employing formal surveys of a broad group of stakeholders and focus groups with key organizations involved in the response to identify insights about managing long duration, moderate acuity public health incidents.

Multiple methods. Gilson and colleagues (2011) recommend the use of multiple research methods within case studies. They suggest that, in a healthcare setting, researchers could conduct two sets of formal interviews with all sampled staff, informal observation and discussion, interviews with patients, and interviews with facility supervisors and area managers at each case study site. Quantitative studies such as analyses of natural experiments or improvement projects can sometimes be embedded into case studies as one of the methods.

Triangulation. Gilson and colleagues (2011) and Yin (2009) both suggest looking for patterns of convergence and divergence by comparing results across multiple sources of evidence (e.g., across interviewees and between interview and other data), between researchers, across methodological approaches, with theory. Stoto, Zhang, and colleagues, for instance, compared multiple data sources to identify a consistent pattern of biases in 2009 H1N1 surveillance data (Stoto, 2012a; Zhang, 2011).

Prolonged engagement with the subject of inquiry, respondent validation, and peer debriefing. As described by Gilson and colleagues, health policy and systems research tends to draw on lengthy and perhaps repeated interviews with respondents, and/or days and weeks of engagement within a case study site (Gilson, 2011). Pawson, Tilley, Greenhalgh, and colleagues make a similar point about evaluations in the realist perspective (Pawson and Tilley, 1997; Greenhalgh, 2009). Rather than maintaining an arm's-length relationship with research subjects, this approach to assessment incorporates reviews of preliminary findings and reports by the PHEP practitioners whose systems are being evaluated. Review of findings and reports by other researchers is also critical.

Clear report of methods of data collection and analysis (audit trail). Gilson and colleagues note the importance of keeping a full record of activities that can be opened to others, and presenting a full account of how methods evolved to the research audience (Gilson, 2011). This can be difficult in studying actual events, which can be long and drawn out, when the focus is naturally on dealing with the public health emergency. Debriefing participants "within the hour," as Weick and Sutcliffe advocate

(Weick and Sutcliffe, 2001), can be difficult in public health emergencies, but the principle of recording the "facts" in real time and saving the analysis for later is important. For example, although not a CDC requirement, the expectation that state and local health department AARs and improvements plans were required to be completed within 60 days of the 2009 H1N1 made it difficult to conduct a thorough analysis of the public health system response (Stoto, 2012b). To address this problem, the National Transportation Safety Board (NTSB) requires a preliminary report within five days of an airplane crash, focusing only on the basic facts, and a factual report with additional information concerning the occurrence within a few months. A final report, which includes a statement of the probable cause, may not be completed for months or until after the investigation has been completed (NTSB, no date).

How well the metrics work: Assessing validity, reliability, practicality, and utility

Since not all proposed measures will necessarily work well in practice, the properties of new and existing PHEP measures must be tested. Indeed, the IOM *Performance Measurement* report calls for a continuous evaluation process and impact assessment of performance measurement initiatives, including funding for associated research activities (IOM, 2007). Consistent with this goal, many of the key performance measurement systems in healthcare have extensive, multi-year development programs in which the uses, measurement domains, and measures and measurement systems are regularly reviewed and updated. This includes the National Committee for Quality Assurance's Healthcare Effectiveness Data and Information Set (HEDIS) (NCQA, no date) and the DHHS family of healthcare consumers' surveys, Consumer Assessment of Healthcare Providers and Systems (CAHPS) (AHRQ, no date).

Such a program, which necessarily involves the measurement system users as well as those being evaluated, will eventually also improve the practicality and utility of the measures. The process should be iterative in the sense that evaluations of the measurement system as implemented will trigger discussions about the ways that the measures have actually been used (why measure); the appropriate domains to be measured as uses and applications shift (what to measure); and new, more valid, reliable, or practical approaches to measurement (how to measure).

Validity is the extent to which assessments really measure the attributes they seek to measure. One aspect is construct or "face" validity, the degree to which the concepts to be measured are translated into specific measures. The validity of measurements based on observations of tabletop exercises, for example, depend on the realism of the scenario, the participation of the decisionmakers who would be involved in a real event, and other factors (Biddinger, 2008). Similarly, the validity of time-to measures based on drills depends on whether the participants have been warned to expect a drill. Face validity is generally assessed by a careful assessment of the biases that are likely to crop up in practice.

For measures based on surveys or other situations where observations from multiple individuals are available, statistical methods from the field of psychometrics assess criterion-related validity, the degree to which the instrument being used correlates with other known and validated measures, either simultaneous (concurrent validity) or future (predictive validity). Validity can also be assessed by examining whether the data structure that emerges from a statistical factor analysis is consistent with what one expects from theory.

Reliability is the extent to which assessments provide consistent measures over time and across measurement units that are not overly influenced by random measurement error due, for example, to sampling individuals to complete a survey. Too much random variation in measures could mean that the measures do not reflect real improvements in the systems being evaluated or that spurious differences over time or between units appear to be real changes or differences. The reliability of measurement systems is assessed by analyzing repeated measurements from the same units, or ensuring that different survey questions in the same domain are consistent with each other. Reliability can be improved by increasing sample size for surveys or averaging independent measures of the same performance.

HSPH CPHP investigators have addressed questions related to the reliability and validity of PHEP measurement tools, focusing in particular on instruments used during discussion-based, tailored, tabletop exercises (Savoia, 2009a). A statistical evaluation of 38 separate PHEP exercises employing realistic scenarios demonstrated the validity and reliability of measures based on participant and external observers that were useful in clarifying public health workers' roles and responsibilities, facilitating knowledge transfer among these individuals and organizations, and identifying specific public health systems-level challenges. Similarly, Savoia and colleagues (2010) analyzed results from two survey instruments developed to assess the performance of Medical Reserve Corps (MRC) volunteers who were able to participate in flu clinics, the specific barriers that prevented some volunteers from participating, and the overall attitudes of those who participated and those who did not. They found the survey instruments to be valid and reliable means to assess the performance and attitudes of MRC volunteers and barriers to their participation.

References

Acosta J, Nelson C, Beckjord EB, *et al*. (2009). A national agenda for public health systems research on emergency preparedness, RAND TR-660. http://www.rand.org/pubs/technical_reports/TR660.html

Agency for Healthcare Research and Quality (AHRQ, no date). CAHPS: Surveys and tools to advance patient-centered care. https://www.cahps.ahrq.gov/default.asp

Aledort JE, Lurie N, Ricci K, *et al*. (2006). Facilitated look-backs: A new quality improvement tool for management of routine annual and pandemic influenza, RAND TR-320. http://www.rand.org/pubs/technical_reports/TR320.html

Andrews FM (1989). Developing indicators of health promotion: Contributions from the social indicators movement. In S.B. Kar (*ed*.), *Health Promotion Indicators and Actions* (pp. 23-49). New York: Springer.

Asch SM, Stoto MA, Mendes M, *et al*. (2005). A review of instruments assessing public health preparedness, *Public Health Reports* 120:532-542.

Association of Public Health Observatories (2008).*The good indicators guide: Understanding how to use and choose indicators*. NHS Institute for Innovation and Improvement. http://www.apho.org.uk/resource/item.aspx?RID=44584

Balicer RD, Barnett DJ, Thompson CB, *et al*. (2010). Characterizing hospital workers' willingness to report to duty in an influenza pandemic through threat- and efficacy-based assessment, *BMC Public Health* 10:436.

Barnett DJ, Balicer RD, Thompson CB, *et al*. (2009). Assessment of local public health workers' willingness to respond to pandemic influenza through application of the extended parallel process model, *PLOS One* 4:e6365.

Barnett DJ, Levine R, Thompson CB, *et al.* (2010). Gauging U.S. emergency medical services workers' willingness to respond to pandemic influenza using a threat- and efficacy-based assessment framework, *PLOS One* 5:e9856.

Beitsch LM, Kodolikar S, Stephans T, *et al*. (2006). A state-based analysis of public health preparedness programs in the United States, *Public Health Reports* 121:737-745.

Biddinger P, Cadigan RO, Auerbach B, *et al*. (2008). Using tabletop exercises to identify systems-level changes for improving preparedness, *Public Health Reports* 123:96-101.

Brand M, Kerby D, Elledge B, *et al*. (2006). A model for assessing public health emergency preparedness competencies and evaluating training based on the local preparedness plan, *Journal of Homeland Security and Emergency Management* 3:1-19.

Centers for Disease Control and Prevention (CDC, 2008). Public Health Emergency Preparedness Cooperative Agreement, Performance Measures Guidance, Budget Period BP9.

Derose SF, Schuster MA, Fielding JE, *et al*. (2002). Public health quality measurement: Concepts and challenges, *Annual Review of Public Health* 23:1-21.

Donabedian A (2005). Evaluating the quality of medical care, *Milbank Quarterly* 83:691-729.

Galarce EM, *et al.* (2011). Socioeconomic status, demographics, beliefs and a (H1N1) vaccine uptake in the United States, *Vaccine* 29:5284-5289.

Gilson L, *et al*. (2011). Building the Field of Health Policy and Systems Research: Social Science Matters, *PLoS Medicine* 8:e1001079.

Greenhalgh T, *et al.* (2009). How Do You Modernize a Health Service? A Realist Evaluation of Whole-Scale Transformation in London, *Milbank Quarterly*, 87:391-416.

Institute of Medicine (IOM, 1988). The Future of Public Health. National Academy Press. http://www.iom.edu/Reports/1988/The-Future-of-Public-Health.aspx

IOM (2003). The Future of the Public's Health in the 21st Century. National Academy Press. http://www.nap.edu/catalog.php?record_id=10548

IOM (2007). Performance Measurement: Accelerating Improvement. National Academies Press. http://www.nap.edu/catalog.php?record_id=11517

IOM (2008). Research Priorities In Emergency Preparedness And Response For Public Health Systems: A Letter Report. National Academy Press. http://www.iom.edu/Activities/Research/PreparednessEMS.aspx

Klaiman T, Kraemer JD, Stoto MA (2011). Variability in school closure decisions in response to 2009 H1N1, *BMC Public Health*, 11:73.

Klaiman, T, O'Connell K, Stoto MA (2012). Local Health Department Public Vaccination Clinic Success During 2009 pH1N1. Forthcoming in *Journal of Public Health Management and Practice*.

Koh HK, Elqura LJ, Judge CM, *et al.* (2008). Regionalization of local public health systems in the era of preparedness, *Annual Review of Public Health* 29:205-218.

Lee BY, Brown ST, Cooley P, *et al.* (2010). Simulating school closure strategies to mitigate an influenza epidemic, *Journal of Public Health Management and Practice* 16:252-261.

Lotstein D, Seid M, Ricci K, *et al.* (2008). Using quality improvement methods to improve public health emergency preparedness: PREPARE for pandemic influenza, *Health Affairs (Web exclusive)* http://content.healthaffairs.org/cgi/content/abstract/hlthaff.27.5.w328

Lovelace K, Bibeau D, Gansneder B, *et al*. (2007). All-hazards preparedness in an era of bioterrorism funding, *Journal of Public Health Management and Practice* 13:465-468.

Lurie N, Wasserman J, Stoto MA, *et al.* (2004). Local variation in public health preparedness: Lessons from California. *Health Affairs Web Exclusive*, http://content.healthaffairs.org/content/early/2004/06/02/hlthaff.w4.341/suppl/DC1

March JG, Sproull LS, and Tamuz M (1991). *Organization Science*, 2:1-13.

Martin LA, Nelson EC, Lloyd RC, Nolan TW (2007). Whole System Measures. IHI Innovation Series white paper. Institute for Healthcare Improvement.

Massachusetts Department of Public Health (MDPH, 2010) Massachusetts H1N1 After-Action Report: H1N1 Pandemic Response – Fall and Winter 2009-10.

National Center for Quality Assurance (NCQA, no date). HEDIS &Quality Measurement. http://www.ncqa.org/tabid/59/Default.aspx

National Quality Forum (NQF, no date a). The ABCs of Measurement. http://www.qualityforum.org/Measuring_Performance/ABCs_of_Measurement.aspx

NQF (no date b). Measure Evaluation Criteria and Guidance Summary Tables. http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=66795

National Transportation Safety Board (NTSB, no date). Aviation Accident/Incident Database. http://www.asias.faa.gov/portal/page/portal/ASIAS_PAGES/ LEARN_ABOUTS/ntsb_la.html

Nelson C, Lurie N, Wasserman, J (2007a). Conceptualizing and defining public health emergency preparedness, *American Journal of Public Health* 97:S9-11.

Nelson C, Lurie N, Wasserman J (2007b). Assessing public health emergency preparedness: Concepts, tools and challenges, *Annual Review of Public Health* 28:1-18.

Nelson C, Chan E Fan C, *et al.* (2009). New Tools for Assessing State and Local Capabilities for Countermeasure Delivery, RAND TR-665. http://www.rand.org/pubs/technical_reports/TR665/

Nelson C, Chan E, Chandra A, *et al.* (2010a). Developing National Standards for Public Health Emergency Preparedness with a Limited Evidence Base, *Disaster Medicine and Public Health Preparedness* 4:285-290.

Nelson C, Parker A, Shelton S, Chan, E (2010b). Answering the call for performance accountability in public health preparedness. Draft manuscript.

Nelson C, Plough A (2011). Managing long duration, moderate acuity public health incidents. Draft manuscript.

Parker A, Nelson C, Shelton S, *et al.* (2009). Measuring crisis decision-making for public health emergencies, RAND TR-712-DHHS. http://www.rand.org/pubs/technical_reports/2009/RAND_TR712.pdf

Pawson R and Tilley N (1997). *Realist Evaluation*. Sage.

Potter MA, Brown ST, Epstein JM, *et al.* (2012). public health system preparedness: a framework for modeling pandemic mitigation strategies. Draft manuscript.

San Francisco Bay Area Advanced Practice Center (2010). The seasonal and pandemic influenza vaccination assessment toolkit. http://sfbayapc.org/document.html?id=64

Savoia E, Testa MA, Biddinger P, *et al.* (2009a). Assessing public health capabilities during emergency preparedness tabletop exercises: Reliability and validity of a measurement tool, *Public Health Reports* 124:139-148.

Savoia E, Massin-Short S, Rodday A, *et al.* (2009b). A literature review of public health systems research in emergency preparedness, *American Journal of Preventive Medicine* 37:150-156.

Savoia E, Massin-Short S, Higdon MA, *et al.* (2010). A toolkit to assess Medical Reserve Corps Units' performance, *Disaster Medicine and Public Health Preparedness* 4:1-7.

Shelton S, Nelson C, McClees A, *et al.* (2011). Building performance-based accountability with a limited empirical evidence base: performance measure development for public health preparedness. Draft manuscript.

Simon H (1981). *The Sciences of the Artificial.* MIT Press.

Stamp JC (1929). Some Economic Factors in Modern Life (King and Son, 1929; p. 258), quoted in Harold Cox (Wikipedia, http://en.wikipedia.org/wiki/Harold_Cox).

Stecher BM, Camm F, Damberg CL (2010). Toward a culture of consequences: performance-based accountability systems for public services, RAND MG-1019. http://www.rand.org/pubs/monographs/2010/RAND_MG1019.pdf

Stoto MA (1992). Public health assessment for the 1990s. *Annual Review of Public Health* 13:59-78.

Stoto MA, Dausey D, Davis L, *et al.* (2005). Learning from experience: The public health response to West Nile Virus, SARS, Monkeypox, and hepatitis A outbreaks in the United States. RAND TR-285. http://www.bvsde.paho.org/bvsacd/cd57/RAND_TR285.pdf.

Stoto MA, Cosler LE (2007). Evaluation. In: Novick L, Mays G, (*eds*). *Public Health Administration: Organization and Strategy for Population-Based Management*, 2nd Edition (pp. 495-544). Jones and Bartlett.

Stoto MA (2008). Regionalization in local public health systems: Variation in rationale, implementation, and impact on public health preparedness. *Public Health Reports* 123: 441-449.

Stoto MA (2012a). How effectively did US public health surveillance systems provide situational awareness during the 2009 pandemic? Forthcoming in *PLoS One*, 2012.

Stoto MA, Nelson C, Higdon MA, *et al.* (2012b). Learning About After Action Reporting from the 2009 H1N1 Pandemic: A Workshop Summary. Draft manuscript.

Trust for America's Health (TFAH, 2011). Ready or not? Protecting the public's health from diseases, disasters and bioterrorism. Annual Report from the Trust for America's Health. http://healthyamericans.org/report/92/

Weick K and Sutcliffe K (2001). *Managing the Unexpected: Assuring High Performance in an Age of Complexity*, Jossey-Bass.

Witte K (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs* 59: 329–349.

Yin R (2009). *Case Study Research: Design and Methods*, Sage.

Zhang Y, May L, Stoto MA (2011). Evaluating syndromic surveillance systems at institutions of higher education (IHEs) during the 2009 H1N1 influenza pandemic. *BMC Public Health*, 11:591.

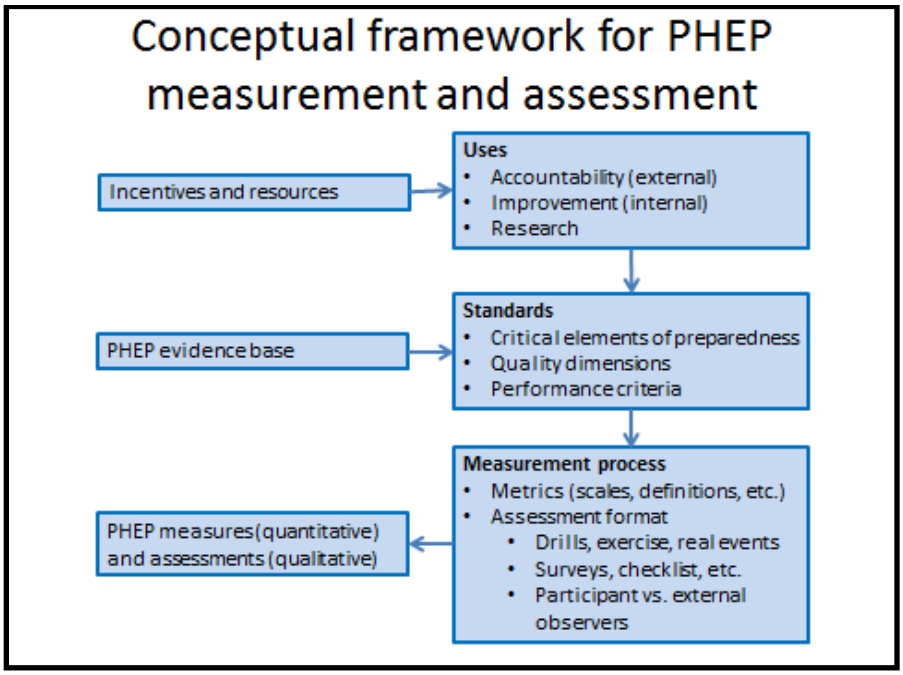

Figure 1.Conceptual framework for assessment. Adapted from Nelson *et al*., 2007b.

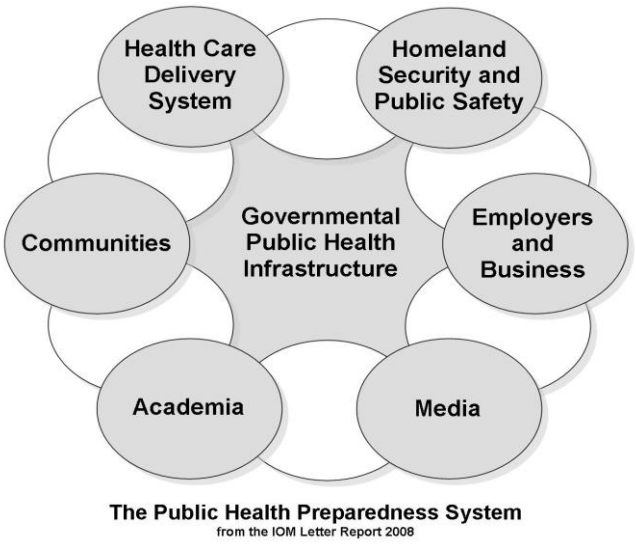

Figure 2.The public health preparedness system. IOM, 2008.

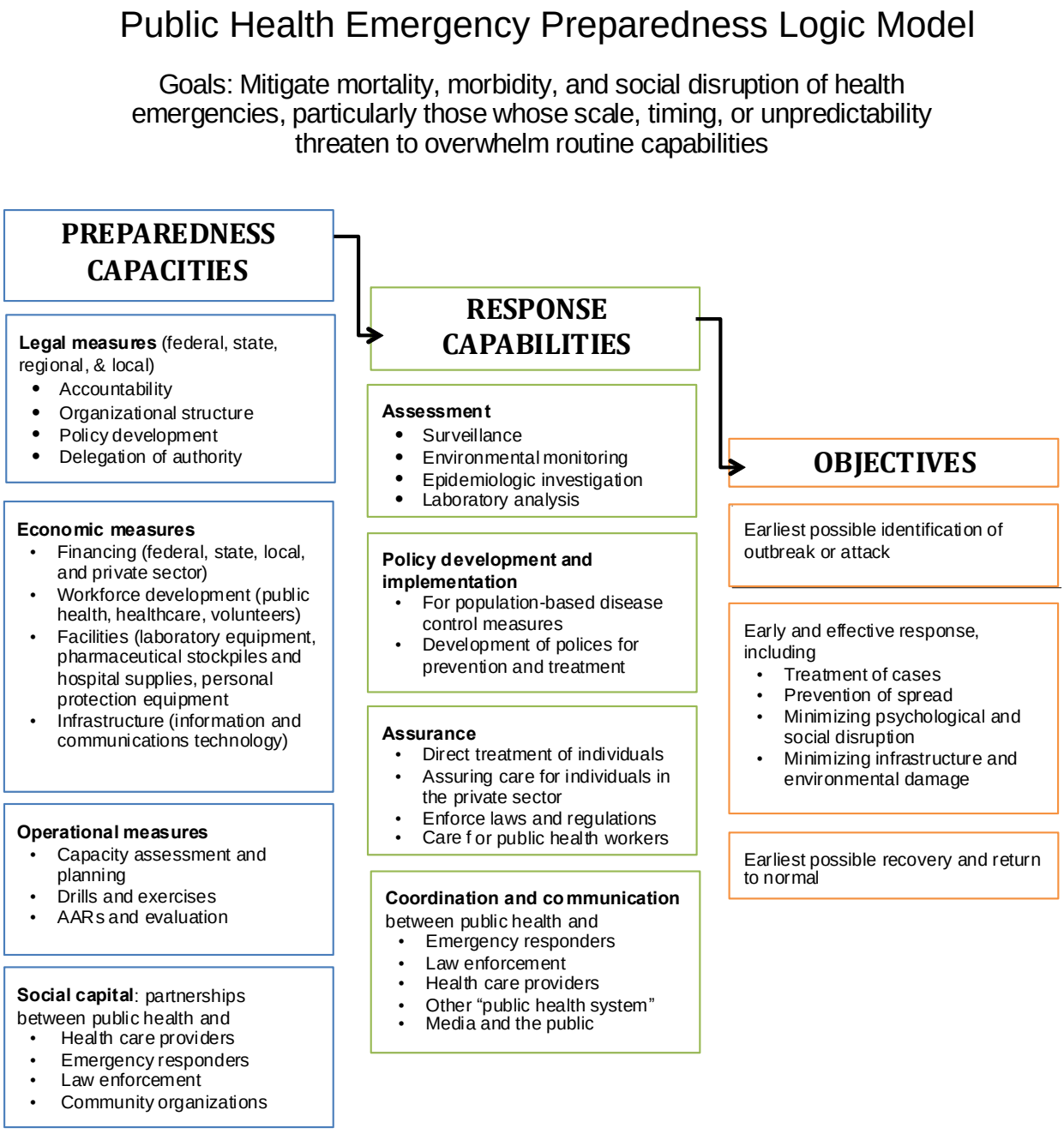

Figure 3.Public health preparedness logic model. Source: Adapted from Stoto *et al*., 2005.

Box 1. Processes for ensuring rigor in case study and qualitative data collection and analysis (Source: Gilson *et al., PLoS Med*, 2011)

- Prolonged engagement with the subject of inquiry. Although ethnographers may spend years in the field, HPSR tends to draw on lengthy and perhaps repeated interviews with respondents, and/or days and weeks of engagement within a case study site
- Use of theory. To guide sample selection, data collection and analysis, and to draw into interpretive analysis
- Case selection. Purposive selection to allow prior theory and initial assumptions to be tested or to examine ‘‘average’’ or unusual experience
- Sampling. Of people, places, times, etc., initially, to include as many as possible of the factors that might influence the behavior of those people central to the topic of focus (subsequently extend in the light of early findings) Gather views from wide range of perspectives and respondents rather than letting one viewpoint dominate
- Multiple methods. For each case study site: two sets of formal interviews with all sampled staff, researcher observation & informal discussion, interviews with patients, and interviews with facility supervisors and area managers
- Triangulation. Looking for patterns of convergence and divergence by comparing results across multiple sources of evidence (e.g., across interviewees, and between interview and other data), between researchers, across methodological approaches, with theory
- Negative case analysis. Looking for evidence that contradicts your explanations and theory, and refining them in response to this evidence
- Peer debriefing and support. Review of findings and reports by other researchers
- Respondent validation (member checking). Review of findings and reports by respondents
- Clear report of methods of data collection and analysis (audit trail). Keeping a full record of activities that can be opened to others and presenting a full account of how methods evolved to the research audience

Box 2. Example measures for the surveillance and epidemiology and disease control and prevention domains for a pandemic influenza scenario

Surveillance and epidemiology domain

- receive and respond to urgent case reports
- investigate and track reported cases
- track information (e.g., newly hospitalized cases, newly quarantined cases)
- laboratory capacity (e.g., rapid identification of unusual influenza strains), including ability to ship specimens to state or CDC lab
- link with and share data among different surveillance systems (e.g., state DOH, CDC, local hospitals)
- step up surveillance capacity in time to initiate containment protocols

Disease control and prevention domain

- knowledge of the legal authorities regarding isolation and quarantine
- procedures to manage isolation and quarantine
- capacity to support people in quarantine
- develop infection control policies and disseminate to hospitals and health care providers
- implement community interventions
- conduct mass screening
- distribute limited medical supplies to priority groups
- control population movement in and out of the community