Assessing Public Health Capabilities During Emergency Preparedness Tabletop Exercises: Reliability and Validity of a Measurement Tool

Elena Savoia, MD, MPH[a]
Marcia A. Testa, PhD, MPH[a,b]
Paul D. Biddinger, MD[a,c]
Rebecca O. Cadigan, ScM[a]
Howard Koh, MD, MPH[a]
Paul Campbell, DSc[a]
Michael A. Stoto, PhD[a,d]

SYNOPSIS

Objectives. Improving the ability of local public health agencies to respond to large-scale emergencies is an ongoing challenge. Tabletop exercises can provide an opportunity for individuals and groups to practice coordination of emergency response and evaluate performance. The purpose of this study was to develop a valid and reliable self-assessment performance measurement tool for tabletop exercise participants.

Methods. The study population comprised 179 public officials who attended three tabletop exercises in Massachusetts and Maine between September 2005 and November 2006. A 42-item questionnaire was developed to assess five public health functional capabilities: (*1*) leadership and management, (*2*) mass casualty care, (*3*) communication, (*4*) disease control and prevention, and (*5*) surveillance and epidemiology. Analyses were undertaken to examine internal consistency, associations among scales, the empirical structure of the items, and inter-rater agreement.

Results. Thirty-seven questions were retained in the final questionnaire and grouped according to the original five domains. Alpha coefficients were 0.81 or higher for all scales. The five-factor solution from the principal components analysis accounted for 60% of the total variance, and the factor structure was consistent with the five domains of the original conceptual model. Inter-rater agreement ranged from good to excellent.

Conclusions. The resulting 37-item performance measurement tool was found to reliably measure public health functional capabilities in a tabletop exercise setting, with preliminary evidence of a factor structure consistent with the original conceptualization and of criterion-related validity.

[a]Center for Public Health Preparedness, Harvard School of Public Health, Boston, MA

[b]Department of Biostatistics, Harvard School of Public Health, Boston, MA

[c]Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA

[d]School of Nursing and Health Studies, Georgetown University, Washington DC

Address correspondence to: Elena Savoia, MD, MPH, Center for Public Health Preparedness, Division of Public Health Practice, Harvard School of Public Health, 677 Huntington Ave., Boston, MA 02115; tel. 617-384-8549; fax 617-495-8543; e-mail <esavoia@hsph.harvard.edu>.

©2009 Association of Schools of Public Health

Several large-scale emergencies, from the 2001 anthrax attacks to the 2005 Gulf Coast hurricanes, have prompted both enhancements to the public health system in responding to emergencies[1] and improved investments in the public health infrastructure at the state and local levels.[2] In this new era, Nelson and colleagues define public health emergency preparedness (PHEP) as "the capability of the public health and health-care 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."[3] This capability involves a coordinated and continuous process of planning and implementation that relies on measuring performance and taking corrective actions.

Because large-scale public health emergencies are infrequent, innovative methods are needed to evaluate PHEP. One possibility is to employ exercises that simulate community response to major emergencies, which familiarize personnel with emergency plans, allow different agencies to practice working together, and identify gaps and shortcomings in emergency planning.[4–6]

Although several instruments for measuring preparedness currently exist, evidence of their validity, reliability, and interpretability is often lacking.[7] To date, only one published study has focused on this important aspect.[8] In addition, current measures are designed to assess quantities of infrastructure elements and variability in capacity. While these have value, also needed are measures of a public health system's capabilities, or its ability to undertake functional or operational actions to effectively identify, characterize, and respond to emergencies. Public health system capabilities include surveillance, epidemiologic investigation, laboratory testing, disease prevention and mitigation, surge capacity for health-care services, risk communication to the public, and the ability to coordinate system response through effective incident management. As proxies for actual emergencies, exercises provide opportunities to identify how various elements of the public health system interact and function to successfully respond, thereby allowing evaluation of system-level preparedness against a known benchmark. In addition, exercises can be used to evaluate the performance of individuals, specific agencies, or an overall (multiagency) system.[9]

In recent years, our group has gained experience in the design, implementation, and evaluation of emergency preparedness tabletop exercises for the public health workforce, with more than 25 exercises conducted throughout New England since 2005. These exercises have been designed and evaluated using a systems-level approach, based on the assumption that the capacity of any organization to respond to a public health threat is influenced by the system in which it is located, just as the preparedness of the overall system is influenced by the strengths and weaknesses of each agency.[9,10]

The aim of this study was to evaluate a performance measurement tool designed to assess public health emergency capabilities, testing its reliability and validity using data obtained both from public officials participating in tabletop exercises and from external evaluators observing the same event. To our knowledge, this is the first study conducted to assess the validity and reliability of a capability measurement tool in the field of PHEP used in the context of a tabletop exercise.

CONCEPTUAL FRAMEWORK AND ITEM-POOL GENERATION

The measurement tool was based upon a public health preparedness logic model, developed by Stoto and colleagues in 2005 to examine the public health response to West Nile virus, severe acute respiratory syndrome, monkeypox, and hepatitis A outbreaks in the U.S.[11] Developed a priori, the model was found to accurately represent the case study findings. Consistent with Nelson and colleagues,[3] the model specifies that the overall goals and objectives of public health preparedness are "to prevent, protect against, quickly respond to, and recover from health emergencies." For tabletop exercises as well as the original case studies, the focus of the model is on functional capabilities—what a public health system needs to be able to perform during an emergency. In the logic model, these capabilities are based upon the three core public health functions: assessment, policy development, and assurance, as described by the Institute of Medicine.[12] Coordination and communication, which include leadership and management, were added because of their clear importance during an emergency response.

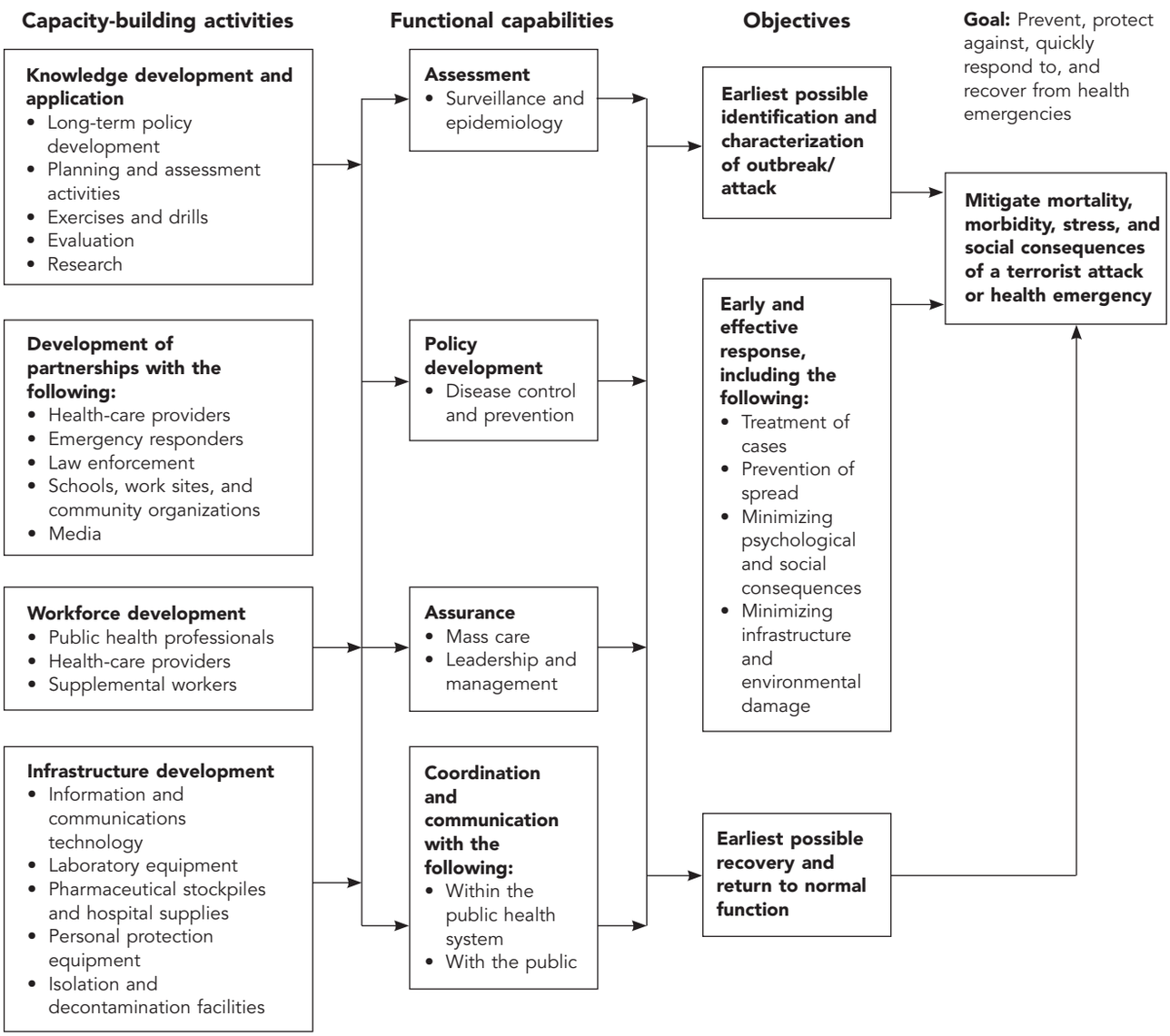
Using the aforementioned logic model and a Modified Delphi technique,[13,14] we convened a panel of 10 senior-level emergency preparedness experts to identify the public health functional capabilities that reasonably could be tested during a tabletop exercise. The panel represented the following areas of expertise: public health practice, hospital preparedness, quantitative methods, epidemiology, biostatistics, surveillance, PHEP assessments, leadership and conflict resolution, and international public health. The panel achieved consensus within the second round of the Modified Delphi process (third draft of the instrument), and

the logic model was adapted to reflect these results (Figure 1). The panel vetted five out of six initially proposed domains: (*1*) leadership and management, (*2*) mass casualty care, (*3*) communication, (*4*) disease control and prevention, and (*5*) surveillance and epidemiology. The sixth domain, law and ethics, was excluded because our tabletop exercises were not designed to trigger significant discussion of these issues. Subsequently, the panel vetted 42 indicators/questions out of 45 and made recommendations concerning items format and questionnaire development. The specific indicators presented to the panel were based on details of the logic model, as well as items from similar checklists for other exercises.[15,16]

DESCRIPTION OF ITEMS IN DEVELOPMENT SAMPLE

We created a 42-item questionnaire. Eleven items addressed the public officials' ability to perform leadership and management tasks. Nine items described the public health system's ability to provide mass care,

Figure 1. Logic model for the assessment of functional capabilities during tabletop exercises[a]



[a]Adapted from: Stoto MA, Dausey D, Davis L, Leuschner K, Lurie N, Myers SS, et al. Learning from experience: the public health response to West Nile virus, SARS, monkeypox, and hepatitis A outbreaks in the United States. Santa Monica (CA): RAND Health; 2005.

including issues related to continuity of operations and the ability to gather and manage physical resources. Communication ability was captured by eight items, five describing the level of communication within the public health system and three addressing the ability of local agencies to keep the public informed. Disease control and prevention were captured by eight items, including ability to conduct mass screening, medical interventions, and law and security issues. Surveillance and epidemiologic ability were addressed by six items, including the ability to implement prompt investigations and to coordinate the response with agencies at the state and/or federal level. For each item, respondents indicated their community's collective ability to perform a specific function using a three-point Likert scale ranging from one (not sufficient) to three (exceeds expectations). We used a three-point Likert scale rather than one with more values because we did not feel that the exercise format supported finer distinctions.

Study participants and setting

The study population comprised 179 public officials representing 59 different municipalities, with population sizes ranging from 603 to 94,304 residents. Each subject attended one of three regional tabletop exercises held in Massachusetts and Maine between September 2005 and November 2006. Participants included local, regional, and state-level professionals from a variety of disciplines such as public health, health care, law enforcement, fire services, emergency medical services, emergency management, and government. Exercises held in Massachusetts were planned in collaboration with the Massachusetts Department of Public Health as well as with selected coalitions of local health departments and boards of health. Exercises held in Maine were planned in collaboration with the Maine Regional Resource Centers for Emergency Preparedness. Exercises were organized by preparedness region in Massachusetts and by county in Maine.

The central goal of each exercise was to provide an opportunity for each region or county to test its community-wide emergency response plan and determine how local government agencies and community groups would work together to respond to a major public health crisis, such as pandemic influenza. Each exercise scenario opened with the announcement of the first human case of avian influenza type A (H5N1) in the U.S., and progressed rapidly to produce widespread infection throughout the state. While the scenario used for each exercise was customized to reflect the unique demographic characteristics, response structure, and assets of that particular region or county, all scenarios spanned a timeline of 14 days (in approximately four to five hours of presentations and discussion), and were designed to test a range of local and regional capabilities: unified command, regional coordination, inter- and intra-agency communications, risk communication to the public, infection control, surge capacity, and mass casualty care. A detailed description of the exercise scenario can be found in the referenced literature.[17]

For each exercise, participants were strategically assigned to interdisciplinary groups of eight to 12 people from the same or neighboring municipalities. Only for municipalities with small population sizes (<2,500) were participants combined with neighboring communities of similar size. At the end of the exercise, each participant was asked to complete a questionnaire to assess his or her community's ability to respond to the emergency described in the scenario, focusing on a set of specific capabilities and indicators. For one of the three exercises, 12 external evaluators were also recruited to independently judge the communities' performances using the same set of indicators. The external evaluators were professionals in the field of emergency preparedness who were considered knowledgeable about emergency preparedness and the plans and organizational structure of the preparedness system tested during the exercise; their years of experience in the field of emergency preparedness ranged from three to 10.

External evaluators observed the communities' performances throughout the exercise and subsequently convened after the exercise to debrief and evaluate capabilities at the community level. Additional scoring guidance was not provided to participants; rather, assessment was based on subjective judgment. However, some guidance was given to the evaluators during the training phase; they were informed in advance about which community they were to observe and the content of the exercise scenario. They were also provided with information regarding the community's demographic characteristics and available resources. In addition, evaluators were trained both by the exercise controller (co-author PDB) on the capabilities expected to be challenged during the event, and by an emergency preparedness evaluation expert (co-author MAS) on their specific role during the exercise, how to use the instrument, and how to identify anchors for each set of indicators, providing them with practical examples.

Assessment of reliability and validity

In this study, we endeavored to assess the reliability, construct validity, and criterion validity of our instrument. In measurement, reliability is an estimate of the

degree to which a scale measures a construct consistently when it is used under the same condition with the same or different subjects. One aspect of reliability is the repeatability of the measurement, and the stability and consistency of the scale over time. Another aspect is the internal consistency of the scale based upon the average intra-item correlations among different items on the same questionnaire. It measures whether several items that propose to measure the same general construct produce similar scores. Validity is the degree to which a test measures what it was designed to measure. Construct validity refers to the correspondence between the concepts (also known as constructs) under investigation and the actual measurements used. It is an iterative process of assessing whether the measures in fact correspond to the concepts that have been identified within a clear conceptual framework. Criterion-related validity is used to demonstrate the accuracy of a measure or procedure by comparing it with another previously validated measure or procedure. Measures must be both reliable and valid, and both properties must be assessed as part of the scale development and validation process.

STATISTICAL METHODS

In the present study, we employed internal consistency as a measure of reliability. Internal consistency of multi-item scales (questions within the same domain) was calculated by means of Cronbach's alpha.[18] Correlations were determined by Pearson's product moment coefficient. The co-linearity of the inter-correlation matrix was tested using Bartlett's Test of Sphericity and by Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy.[19] The empirical structure of the scales was determined by principal components analysis. The Kaiser-Guttman rule (eigenvalue >1) and the corresponding screen plot results were used to determine the number of factors to be retained. Factors were rotated using varimax rotation. The factor scores were computed as the linear combination of the product of the standardized score and its factor loading.[20]

We also measured validity by comparing participants' responses to the observations of external evaluations. Participants' responses were correlated to those given by the external evaluators using intra-class correlation coefficients (ICC). Mean values by scale for both sources of rating (participants and external evaluators) were graphically displayed. The statistical analysis was performed using the statistical package SPSS version 17.0.[21]

RESULTS

Characteristics of respondents/communities and analysis of missing values

We collected a total of 179 questionnaires during the course of the three exercises. Of the total respondents, 102 were from Maine (57%) and 77 were from Massachusetts (43%). Participants in this sample represented 15 job categories, with highest representation in three categories: public health workers (23%), health-care professionals (21%), and firefighters (12%). Analysis of missing items and not applicable (NA) responses revealed a pattern across exercises. We noted a higher number of missing values (a mean of 30%) and NA responses (a mean of 10%) from questionnaires returned by participants working in rural areas as compared with those from participants working in urban areas. One of the exercises was held in a particularly rural area of Maine that lacks a formal public health infrastructure and includes limited health-care resources. The mean percentage of NA responses per item for this region was 13.8 (standard deviation [SD]=9.3), while in a less rural geographical area it was 5.8 (SD=4.8).

A pattern was also observed in the type of questions that respondents did not answer or found NA. Across all three exercises, the highest number of NA responses was given to a particular question (surveillance and epidemiology domain, question 35) related to laboratory capability levels (i.e., rapid identification of unusual influenza strains), including the ability to ship specimens to state or federal laboratories. The highest number of missing values was related to the following three questions: ability to provide mortuary services (mass care domain, question 19), ability to activate the Strategic National Stockpile (mass care domain, question 15), and ability to step up surveillance capacity in time to initiate containment protocols (surveillance and epidemiology domain, question 37). When performing the factor analysis, we excluded all NA and missing responses. For item imputation, we replaced the missing items with the mean of the non-missing items in the scale if at least 50% of the items were non-missing. Of the 179 returned questionnaires, 126 were considered suitable for statistical analysis.

End users' feedback

The questionnaire also included an open-ended question asking participants and evaluators to add comments on how to improve the instrument. The majority of the comments focused on how to implement the questionnaire and how to train evaluators, rather than on the specific content of the items.

Specifically, evaluators requested more training on how to determine the anchors of their responses and how to translate their subjective judgments into numbers, and several requested additional time to complete the questionnaire. For these reasons, we let evaluators return the questionnaire the day after the exercise, and included a separate form with a set of open-ended questions where they could describe in more detail the community's response in each of the five domains.

Empirical scale development

Data derived from 126 subjects who completed the instrument were determined suitable for principal components factor analysis by inspection of the anti-image covariance matrix, KMO measure of sampling adequacy, and Bartlett Test of Sphericity. The KMO value was 0.75 (satisfactory). The Bartlett Test of Sphericity, which tests for the presence of correlations among the variables, was significant ($p<0.001$), indicating that the correlation matrix was consistent with the hypothesis that the measures were not independent. A five-factor solution (which accounted for 60% of the total variance) was found to be parsimonious, had a simple structure, and could be meaningfully interpreted. Items with factor loadings >0.45 were used to define the factors.

Of the original 42 items, 37 met the criteria. Abbreviated items and their corresponding factor loadings are presented in the Table. Full-item wording and scaling is given in Figure 2. Five items were dropped because of redundancy or because of unclear wording. Two items, both hypothesized to load on the communication domain, were found to fit better with the leadership and management (question 11) and disease control (question 31) domains. As this result could be meaningfully interpreted by the factor structure, the two items were kept in the final instrument. The first factor consisted of 11 items and accounted for 39% of the total variance. All 11 items were related to leadership and management. The second factor consisted of eight items and accounted for 8% of the variance. All eight items addressed mass care capabilities. The third factor accounted for 5% of the variance and comprised four items describing the agencies' ability to maintain adequate communication within the public health system, including exchange of information among policy makers, health-care providers, and public health agencies. The fourth factor accounted for 4% of the variance and comprised eight items describing the ability to address disease control and prevention issues. The fifth factor accounted for 4% of the variance, with six items addressing surveillance and epidemiology capabilities.

Internal consistency

Cronbach's coefficient was calculated for the five summary scales based on the factor analysis, as well as for the total summative score based on all 37 items. The overall measure of internal consistency for the summative score was 0.95. All scale coefficients were 0.81 or higher: leadership and management (0.93), mass care (0.88), communication within the public health system (0.86), disease control and prevention (0.87), and surveillance and epidemiology (0.81). These results indicate that for each of the five domains, questions within each domain were correlated with one another and had robust internal consistency.

Domain structure

In the conceptual model, we hypothesized that all five domains would be positively correlated and that correlations between comparable domains would be stronger than between less comparable domains. The leadership and management domain was expected to have medium or large correlations with the mass care, communication, and disease control domains, and a small correlation with the surveillance and epidemiology domain. Similarly, the surveillance and epidemiology domain was expected to have a larger correlation with the disease control domain relative to the others. The strength of the correlation was determined by Cohen's (1992) criteria, where large correlations are described as being $r \geq 0.50$, medium correlations range from $r=0.30$–0.49, and small correlations range from $r=0.10$–0.29. The results supported the hypothesized structure: the five domains were positively and significantly correlated to one another.

We found moderate correlations between the leadership domain and the mass care ($r=0.67$, $p<0.001$), disease control ($r=0.68$, $p<0.001$), and communication domains ($r=0.66$, $p<0.001$), and a weaker correlation between the leadership domain and the surveillance and epidemiology domain ($r=0.33$, $p<0.001$). In addition, as hypothesized, among all possible pairs of correlations with the surveillance and epidemiology domain, the correlation with the disease control domain was the strongest ($r=0.45$, $p<0.001$).

Inter-rater agreement and criterion-related validity

In the field of organizational performance assessments, the use of trained evaluators is considered to be the best available evaluation approach in terms of perceived objectivity and legitimation for accountability purposes.[22,23] In the field of emergency preparedness, the use of external evaluators has been proposed as an evaluation method to assess organizational and system capabilities. For the purpose of this study,

Table. Questionnaire factor loading

Leadership and management		*Factor loading*
Question 1.	Identify activities that will be performed at a state, local, or coordinated level.	0.63
Question 2.	Interact with local, state, and federal officials with regard to the delegation of legal and law enforcement responsibilities.	0.54
Question 3.	Identify the authority for declaring a public health emergency.	0.54
Question 4.	Gather resources in support of implementing action.	0.71
Question 5.	Assess and manage local resources.	0.73
Question 6.	Address and respond to cross-jurisdictional needs.	0.64
Question 7.	Assist special-needs populations.	0.62
Question 8.	Respond flexibly, in proportion to the magnitude and severity of the scenario and available resources.	0.78
Question 9.	Activate psychosocial needs and activate appropriate services.	0.66
Question 10.	Integrate community-based organizations in the response.	0.61
Question 11.	Communicate with the general public about up-to-date outbreak information, disease control requirements, individual risk reduction, and when and where to seek medical care.	0.52
Mass care		
Question 12.	Ensure continuity of health-care operations.	0.67
Question 13.	Assure health-care for all cases that meets relevant standards of care.	0.70
Question 14.	Protect health-care workers with personal protective equipment and infection control practices.	0.71
Question 15.	Activate and use the Strategic National Stockpile.	0.57
Question 16.	Prioritize the use of limited medical supplies.	0.59
Question 17.	Provide security within health-care facilities and at mass point-of-dispensing sites.	0.57
Question 18.	Coordinate medical reserve.	0.45
Question 19.	Make provisions of mortuary services.	0.63
Communication within the public health system		
Question 20.	Provide current information (i.e., newly hospitalized cases, newly quarantined cases) to policy makers.	0.67
Question 21.	Disseminate infection control policies to hospitals and health-care providers.	0.65
Question 22.	Communicate with public health agencies in neighboring communities and the state.	0.67
Question 23.	Communicate within the local public health system (including other government agencies).	0.68
Disease control and prevention		
Question 24.	Knowledge of the legal authorities regarding isolation and quarantines.	0.50
Question 25.	Availability of procedures to manage isolation and quarantine.	0.55
Question 26.	Capacity to support people in quarantine (e.g., preidentified sites, support for home quarantine).	0.56
Question 27.	Implement community interventions such as school closings.	0.48
Question 28.	Conduct mass screening.	0.54
Question 29.	Distribute limited medical supplies (including vaccines) to priority groups.	0.57
Question 30.	Control population movement in and out of the community.	0.61
Question 31.	Communicate with the public to minimize fear.	0.58
Surveillance and epidemiology		
Question 32.	Receive and respond to urgent case reports.	0.54
Question 33.	Investigate and track reported cases.	0.73
Question 34.	Track information (i.e., newly hospitalized cases, newly quarantined cases) for policy makers.	0.81
Question 35.	Monitor laboratory capacity (i.e., rapid identification of unusual influenza strains), including ability to ship specimens to state or CDC lab.	0.65
Question 36.	Link with and share data among different surveillance systems (e.g., state department of health, CDC, other communities and states, local hospitals).	0.66
Question 37.	Step up surveillance capacity in time to initiate containment protocols.	0.63

CDC = Centers for Disease Control and Prevention

we assumed that the external evaluator assessments qualified as the gold standard. We performed group-level correlations between self and evaluator ratings to test inter-rater agreement, and we considered high agreement as evidence of measurement reliability and criterion validity.

The level of agreement ranged from good to excellent depending on the domain: leadership and management (ICC=0.56), mass care (ICC=0.67), communication within public health (ICC=0.91), disease control and prevention (ICC=0.89), and surveillance and epidemiology (ICC=0.72). Furthermore, the scores reported by external evaluators were on average lower than those reported by the participants

(Figure 3). This pattern is consistent with the literature and further supports the concurrent validity of the instrument.[22,23]

However, for four out of eight items describing mass care capabilities (questions 12–15), there was an inverse relationship such that the score was higher when reported by the external evaluator. We believe this may due to a "table factor," an "evaluator factor," or both. During that specific exercise, representatives from hospitals were grouped together at a separate table, rather than grouped with representatives of their community. Such grouping facilitated coordination among hospitals, and mass care capabilities were perceived as high by the evaluator who was observing that particular table. In addition, it is possible that this result was attributable to an evaluator factor related to lack of evaluation training in that particular area. A further analysis showed that the score given by this evaluator was an outlier that raised the overall mean score of the external evaluators for these four items. These results are displayed in Figure 3.

Figure 2. Avian influenza tabletop exercise evaluation form

In light of your experiences during today's tabletop exercise, please rate your community's ability to perform the tasks listed below, using the following scoring system: 1 (not sufficient), 2 (sufficient), 3 (exceeds expectations), or NA (not applicable).

Your table number and/or town name:
Your professional role:
Your organization:

Leadership and Management

Please rate your community's ability to respond using the following scoring system:
1 = not sufficient
2 = sufficient
3 = exceeds expectations
NA = not applicable

Indicators	*Score*			
1) Identify activities that will be performed at a state, local, or coordinated level.	1	2	3	NA
2) Interact with local, state, and federal officials with regard to the delegation of legal and law enforcement responsibilities.	1	2	3	NA
3) Identify the authority for declaring a public health emergency.	1	2	3	NA
4) Gather resources in support of implementing action.	1	2	3	NA
5) Assess and manage local resources.	1	2	3	NA
6) Address and respond to cross-jurisdictional needs.	1	2	3	NA
7) Assist special-needs populations.	1	2	3	NA
8) Respond flexibly, in proportion to the magnitude and severity of the scenario and available resources.	1	2	3	NA
9) Anticipate psychosocial needs and activate appropriate services.	1	2	3	NA
10) Integrate community-based organizations in the response.	1	2	3	NA
11) Communicate with the general public about up-to-date outbreak information, disease control requirements, individual risk reduction, and when and where to seek medical care.	1	2	3	NA

Mass Care

Indicators	*Score*			
12) Ensure continuity of health-care operations.	1	2	3	NA
13) Assure health care for all cases that meets relevant standards of care.	1	2	3	NA
14) Protect health-care workers with personal protective equipment and infection control practices.	1	2	3	NA
15) Activate and use the Strategic National Stockpile.	1	2	3	NA
16) Prioritize the use of limited medical supplies.	1	2	3	NA
17) Provide security within health-care facilities and at mass point-of-dispensing sites.	1	2	3	NA
18) Coordinate medical reserve.	1	2	3	NA
19) Make provisions of mortuary services.	1	2	3	NA

continued on p. 146

Figure 2 (continued). Avian influenza tabletop exercise evaluation form

Communication Within the Public Health System

Indicators	*Score*			
20) Provide current information (i.e., newly hospitalized cases) to policy makers.	1	2	3	NA
21) Disseminate infection control policies to hospitals and health-care providers.	1	2	3	NA
22) Communicate with public health agencies in neighboring communities and the state.	1	2	3	NA
23) Communicate within the local public health system (including other government agencies).	1	2	3	NA

Disease Control and Prevention

Indicators	*Score*			
24) Knowledge of the legal authorities regarding isolation and quarantine.	1	2	3	NA
25) Availability of procedures to manage isolation and quarantine.	1	2	3	NA
26) Capacity to support people in quarantine (e.g., preidentified sites, support for home quarantine).	1	2	3	NA
27) Implement community interventions such as school closing.	1	2	3	NA
28) Conduct mass screening.	1	2	3	NA
29) Distribute limited medical supplies (including vaccines) to priority groups.	1	2	3	NA
30) Control population movement in and out of the community.	1	2	3	NA
31) Communicate with the public to minimize fear.	1	2	3	NA

Surveillance

Indicators	*Score*			
32) Receive and respond to urgent case reports.	1	2	3	NA
33) Investigate and track reported cases.	1	2	3	NA
34) Track information (i.e., newly hospitalized cases, newly quarantined cases) for policy makers.	1	2	3	NA
35) Laboratory capacity (i.e., rapid identification of unusual influenza strains) including ability to ship specimens to state or CDC lab.	1	2	3	NA
36) Link with and share data among different surveillance systems (e.g., state department of health, CDC, other communities and states, local hospitals).	1	2	3	NA
37) Step up surveillance capacity in time to initiate containment protocols.	1	2	3	NA

CDC = Centers for Disease Control and Prevention

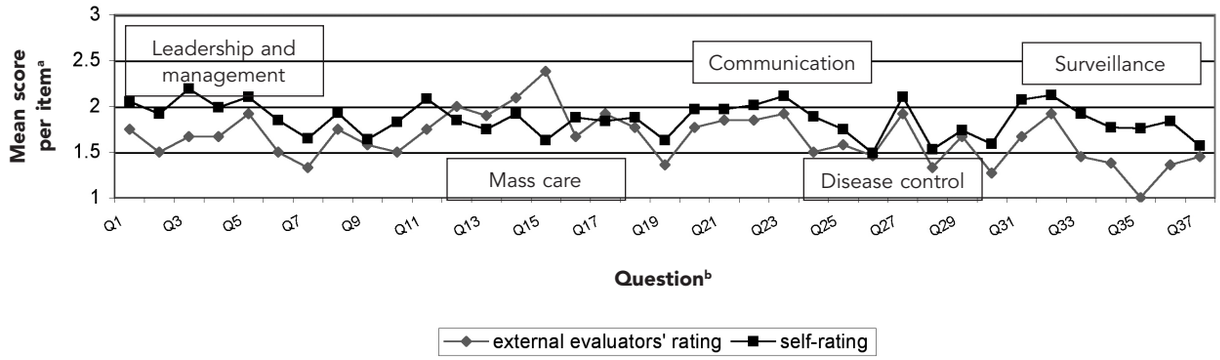
DISCUSSION

While the gold standard of evaluation research involves ascertainment of outcomes, in disaster planning such outcomes are rare. Hence, the most feasible methodological approach is to identify potential predictors or proxies of outcomes and to implement interventions that will improve the outcome measures of those predictors. Tabletop exercises that mimic emergencies are an innovative method that can be used to approximate this ideal by means of assessing public health emergency response capabilities. While it is important to recognize that exercise outcomes are influenced by several factors, including the exercise design; the ability of the controller and facilitators to conduct the exercise; and by the professional role, level of participation, training, and experience of the participants, they are still considered a useful tool to guide improvement efforts and ensure accountability for public investments. Furthermore, while identifying the best tools to measure the quality of performance in tabletop exercises is an evolving science,[24] self-assessment can be one useful and cost-effective measure of performance.[22,23]

To our knowledge, our study is the first to assess the validity and reliability of a capability measurement tool in the field of PHEP used in the context of a tabletop exercise. We assessed public health officials' perceptions of the performance level of their communities' responses to a simulated influenza pandemic and compared them with assessments of external evaluators. Our instrument showed strong evidence of reliability, and a factor structure consistent with the original conceptualization. Only two items out of 37 did not load on the hypothesized domain, both related to risk communication capabilities that instead, and perhaps not surprisingly, fit better with the leadership and management and disease control domains.

It should be noted, however, that the reliability of

Figure 3. External evaluators' rating vs. self-rating



[a]Using a Likert scale, where 1=insufficient and 3=exceeds expectations.

[b]Questions pertained to different topics: leadership (1–11), mass care (12–19), communication within public health system (20–23), disease control (24–31), and surveillance and epidemiology (32–37).

a measure is linked to the population to which one wants to apply the measure. There is no reliability of a test, unqualified; rather, the coefficient has meaning only when applied to specific populations. Moreover, validity is strongly related to the context in which the instrument is implemented, which in this particular case refers to the way the exercises were designed, the discussion was facilitated, and responses were collected. For all these reasons, the reliability and validity of our instrument are sample and context-specific and should be retested if the instrument is used in a setting with a study population that possesses characteristics distinctly different from the population used in this study.

Our results also highlighted the need to adapt performance measurement instruments to reflect differences due to the presentation, test setting, and environment. As noted previously, in an exercise performed in a rural area of Maine, the number of missing values and NA responses was approximately four times the values obtained in the other, more urban geographical areas. We believe this reflected the lack of a formal local public health infrastructure and the presence of relatively few health-care facilities in rural Maine.

CONCLUSIONS

We developed a measurement tool consisting of a 37-item questionnaire and found it to be a reliable measure of public health functional capabilities in tabletop exercises, with preliminary evidence of construct and criterion-related validity. The items clustered in five scales representing the core functions of public health as applied to emergencies: (*1*) assessment, especially surveillance and epidemiology; (*2*) policy development, especially with regard to disease control and prevention; (*3*) assurance, especially the provision of mass care; (*4*) leadership and management; and (*5*) communication within the public health system.

The tool can be implemented during tabletop exercises and completed by participants and external evaluators to successfully identify public health capabilities in need of further improvement. While additional work is needed to continue advancing the field of measuring PHEP, we believe that this tool could serve as a valuable and cost-effective resource to measure a community's emergency planning progress and to determine how further resources might most efficiently be employed to help strengthen the system.

The work of the Harvard School of Public Health Center for Public Health Preparedness is supported under a cooperative agreement with the Centers for Disease Control and Prevention (CDC) grant number U90/CCU124242-03/04.

The discussion and conclusions in this article are those of the authors and do not necessarily represent the views of CDC or the U.S. Department of Health and Human Services.

REFERENCES

1. Lurie N, Wasserman J, Nelson CD. Public health preparedness: evolution or revolution? Health Aff (Millwood) 2006;25:935-45.
2. Department of Health and Human Services (US). HHS announces $896.7 million in funding to states for public health preparedness and emergency response [press release]; 2007 Jul 17 [cited 2008 Sep 2]. Available from: URL: http://sev.prnewswire.com/health-care-hospitals/20070717/DCTU11717072007-1.html
3. Nelson C, Lurie N, Wasserman J, Zakowski S. Conceptualizing and defining public health preparedness. Am J Public Health 2007;97 Suppl 1:S9-11.
4. Stoto MA, Cosler LE. Evaluation. In: Novick LF, Morrow CB, Mays GP, editors. Public health administration: principles for population-based management. 2nd ed. Sudbury (MA): Jones and Bartlett Publishers; 2008. p. 495-544.

5. Dausey DJ, Buehler JW, Lurie N. Designing and conducting tabletop exercises to assess public health preparedness for manmade and naturally occurring biological threats. BMC Public Health 2007;7:92.
6. Biddinger PD, Cadigan RO, Auerbach B, Burnstein J, Savoia E, Stoto MA, et al. Using exercises to identify systems-level preparedness challenges. Public Health Rep 2008;123:96-101.
7. Asch SM, Stoto M, Mendes M, Valdez RB, Gallagher ME, Halverson P, et al. A review of instruments assessing public health preparedness. Public Health Rep 2005;120:532-42.
8. Costich JF, Scutchfield FD. Public health preparedness and response capacity inventory validity study. J Public Health Manag Pract 2004;10:225-33.
9. Moore S, Mawji A, Shiell A, Noseworthy T. Public health preparedness: a systems-level approach. J Epidemiol Community Health 2007;61:282-6.
10. Braun BI, Wineman NV, Finn NL, Barbera JA, Schmaltz SP, Loeab JM. Integrating hospitals into community emergency preparedness planning. Ann Intern Med 2006;144:799-811.
11. Stoto MA, Dausey D, Davis L, Leuschner K, Lurie N, Myers SS, et al. Learning from experience: the public health response to West Nile virus, SARS, monkeypox, and hepatitis A outbreaks in the United States. Santa Monica (CA): RAND Health; 2005.
12. Institute of Medicine. The future of public health. Washington: National Academy Press; 1988.
13. Linstone HA, Turoff M. The Delphi Method: techniques and applications. 2002 [cited 2008 Mar 18]. Available from: URL: http://is.njit.edu/pubs/delphibook
14. Custer RL, Scarcella JA, Stewart BR. The Modified Delphi technique—a rotational modification. J Vocational Technical Ed 1999;15:50-8.
15. Lurie N, Wasserman J, Stoto M, Myers S, Namkung P, Fielding J, et al. Local variation in public health preparedness: lessons from California. Health Aff (Millwood) 2004;Suppl Web Exclusives: W4-341-53.
16. Dausey DJ, Aledort JE, Lurie N. Tabletop exercise for pandemic influenza preparedness in local public health agencies. Santa Monica (CA): RAND Health TR319; 2006. Also available from: URL: http://www.rand.org/pubs/technical_reports/TR319 [cited 2008 Sep 2].
17. Harvard School of Public Health, Center for Public Health Preparedness. Tabletop exercise no. 1: avian/pandemic influenza [cited 2008 Sep 2]. Available from: URL: http://www.hsph.harvard.edu/hcphp/files/HSPH-CPHP_Avian_Pandemic_Influenza_Tabletop.pdf
18. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psykometrika 1951;16:297-334.
19. Hair JF, Anderson RE, Tatham RL, Black WC. Multivariate data analysis: with readings. 4th ed. Englewood Cliffs (NJ): Prentice-Hall; 1995.
20. Dunteman GH. Principal components analysis. Newbury Park (CA): Sage Publications; 1989.
21. SPSS, Inc. SPSS: Version 17.0 for Windows. Chicago: SPSS, Inc.; 2002.
22. Conley-Tyler M. A fundamental choice: internal or external evaluation? Evaluation Journal of Australasia 2005;4:3-11.
23. Dunning D, Heath C, Suls JM. Flawed self-assessment: implications for health, education, and the workplace. Psychological Science in the Public Interest 2004;5:69-106.
24. Dausey DJ, Chandra A, Schaefer AG, Bahney B, Haviland A, Zakowski S, et al. Measuring the performance of telephone-based disease surveillance systems in local health departments. Am J Public Health 2008;98:1706-11.