Novel Process for Developing Metrics That Measure What Police Do

Bryan Vila, Ph.D.  
Dept. of Criminal Justice  
Washington State University  
Spokane, WA  
vila@wsu.edu

Stephen M. James, Ph.D.  
Elson S. Floyd College of Medicine  
Washington State University  
Spokane, WA  
stevejames@wsu.edu

Lois James, Ph.D.  
College of Nursing  
Washington State University  
Spokane, WA  
lois_james@wsu.edu

ABSTRACT

Objectives: Develop and test a technique for objectively measuring individual police officer performance in complex, dynamic encounters with the public, while controlling for situational and environmental difficulty. Assess the technique’s ability to rapidly prototype metrics for measuring officer behavior (rather than outcomes) in different types of encounters in order to evaluate the impact of police policies, practices, and training on performance.

Methods: Metric sets were developed separately for three types of encounters, deadly force judgment and decision making (DFJDM), cross-cultural tactical social interaction (TSI), and crisis intervention (CIT). In each, “reverse concept mapping” was used with a diverse focus group of “true experts” to deconstruct their implicit concepts and derive important performance variables. Variables then were scaled to create metrics for situational difficulty and officer performance. Thurstone scaling was then used with hundreds of diverse expert trainers to create interval-level scales. Metric utility was tested during three different experimental studies.

Results: Three separate metric sets were successfully developed for DFJDM, TSI, and CIT encounters. Collectively, they have been used successfully to control for situational difficulty, establish performance criteria and behavioral objectives, develop high-definition training scenarios, and implement training curricula.

Conclusions: This novel combination of metric development techniques provides a practical and efficient method for rapidly developing metrics that measure micro-level performance during police encounters with the public with interval-level precision. These metrics enable objective and grounded measurement of what officers do as they exercise discretion, engage people, and build or erode public perceptions of legitimacy.

Keywords: accountability, evaluation, metrics, decision making.

ABOUT THE AUTHORS

Bryan Vila, Ph.D., is a professor emeritus at WSU who pioneered the study of police fatigue and leads this area of research internationally. He served as Chief of the Crime Control and Prevention Research Division at the National Institute of Justice. Dr. Vila also has 17 years of experience as a law enforcement officer, including six years as a police chief in the emerging island nations of Micronesia.

Stephen M. James, Ph.D., is an assistant research professor in the Elson S. Floyd College of Medicine at WSU. His research focus includes the interaction between law, policy, training and practice relating to operational performance for military and law enforcement personnel. Prior to becoming an academic, Dr. James spent more than 20 years in the British infantry as a soldier and officer; serving in Cyprus, the former Yugoslavia, Northern Ireland, and Afghanistan.

Lois James, Ph.D., is an assistant professor in the WSU College of Nursing. Dr. James’s simulated research on the impact of suspect race on decisions to shoot has significantly advanced what is known about how suspect race influences police officers during deadly encounters. She is a member of the International Association of Chiefs of Police (IACP) Research Advisory Council.

Acknowledgements
The research presented in this paper has been supported by federal grants from the US DOJ NIJ (grant no. 2008IJCX0015), US DOD DARPA (grant no. W911NF-12-0039), US DOD ONR (grant nos. N000141110185 & N000141512470), and US DOD DPSI via NAVAIR (subcontract no. CBSC-061615-1 on prime contract no. N61340-11-D-0003).
**Novel Process for Developing Metrics That Measure What Police Do**

Bryan Vila, Ph.D.  
Dept. of Criminal Justice  
Washington State University  
Spokane, WA  
vila@wsu.edu

Stephen M. James, Ph.D.  
Elson S. Floyd College of Medicine  
Washington State University  
Spokane, WA  
stevejames@wsu.edu

Lois James, Ph.D.  
College of Nursing  
Washington State University  
Spokane, WA  
lois_james@wsu.edu

**INTRODUCTION**

If the past few years have taught us anything, it’s that the most important aspect of police performance is what individual officers do on the streets. Among the many things that cops do, it’s difficult to think of an area in the study of police performance in greater need of more accurate measurement than how individual officers perform during encounters with the public. Much of the research reported in this paper focuses on encounters that tend to have a higher potential for turning deadly. However, mundane police encounters that influence public perceptions of the police are also critical. As Mastrofski put it, “One cannot overemphasize the importance of doing more to measure the discretion exercised by street-level police officers in deciding when and where to mobilize to do something” (2002: 113). Our research measures the micro-level details of what police do during encounters so that we can assess the string of behaviors that constitute performance and affect the probability of a desirable outcome.

On the streets, most police performance measures are post-hoc, they focus on outcomes rather than what police do or don’t do to achieve those outcomes. Even officer training for encounters with the public usually relies on coarse performance measures (with the obvious exception of skills such as marksmanship). It focuses on complex concepts such as “situational awareness” or “command presence” that generally are measured subjectively by trainers using ordinal ranking or categorical heuristics. Officer performance—and the public’s perceptions of it—directly affect justice on the streets and in the police workplace. It is critical that it be measured in as precise and valid manner as possible.

The novel metric development process reported here addresses this critical need by combining two well-established research techniques, concept mapping (e.g., Novak, J. D. & Gowin, D. B., 1984; Kane and Trochim, 2007) and Thurstone scaling (Thurstone and Chave, 1929; Thurstone, 1959). We used this novel process to identify concrete performance variables associated with police encounters with the public and establish interval-level measures for them, as well as the relative difficulty of the situational environment in which an encounter occurs. Our implementations of this novel combination of techniques have demonstrated that it is a practical, efficient, and cost-effective method for developing metrics that measure what matters with much greater precision than has previously been possible. These improvements are critically important because many of the most divisive and corrosive issues in contemporary American society are anchored in police interactions with the public. Public perceptions of the fairness, respect and legitimacy of what police officers do in encounters has a strong impact on their “compliance and cooperation” with officers (Mazerolle et al., 2013). People care as much—or more—about how they are treated by the police than about changes in crime rates (e.g., Tyler and Huo, 2002, Tyler et al., 2015).

**Dynamic Social Encounters**

How an officer behaves toward people each time he or she encounters them generates a cascade of responses, counter responses, and interactions with active participants and bystanders (Goffman, 1969). This dynamic system of interactions evolves as a network of interactions surges and wanes, each action spawning others, reinforcing some possibilities and missing or countering others. Each actor in this intimate social system tends to try—with more or less success—to assess the probable consequences of the actions they use to influence others in the encounter and guide its course toward a desired outcome (Bakeman & Gottman, 1986).

Of course the physical environment also shapes these social interactions. For example, objects limit fields of view, provide protection and channel movement. Light and sound levels constrain what can be perceived, and how much
time and cognitive effort are required to interpret a perception. Terrain, weather, and the positions of people also affect what can be perceived, options for action, and the consequences of choosing one option over another. This means one must consider the relative situational difficulty of the environment in which encounters between police and the public unfold when evaluating officer performance.

In addition to the challenges presented by situational difficulty, the dynamic forces at play in social encounters become increasingly difficult to understand, control, and predict as they become more situationally and socially interconnected, or as the pace of interactions accelerates (see Klinger, 2004, 2005; Perrow, 1984). This increase in turbulence (McCann & Selsky, 1984) raises the probability of an unforeseen or unforeseeable catastrophe. Catastrophic outcomes in social encounters at the micro level become unforeseeable when the increased volume, speed and coherence of information flowing through the social system overwhelms the cognitive and perceptual abilities that actors in the encounter need in order to assess what’s going to happen next and how best to respond (see Eubank & Farmer, 1990; Holland, 1992:184-185; Mitchell, 2009; Vila, 2010; and Eagleman, 2011 for detailed explanations of these dynamics).

The real-world implication of these dynamics is that encounters between the police and the public are probabilistic. Tragic outcomes can occur despite the best intentions of actors in the encounter—and, conversely, good outcomes can occur despite their worst intentions. Ideally, then, in order to assess the propriety and overall justness of officers’ actions we need to assess what they do in an encounter and the relative difficulty of the situational environment, not just how things turn out. And we need to employ the best quality metrics possible (i.e., those with the greatest external validity and precision) in order to perform the most powerful and definitive statistical analyses possible.

Naturalistic Decision Making

One of the most promising approaches for understanding how people make decisions in situations similar to those faced by police in dynamic social encounters is naturalistic decision making (NDM), which was developed in the late 1980s. In the sorts of complex, dynamic, fast-paced, and low information situations that often characterize police encounters with the public, officers often rely on “expertise”—skilled intuition based on a synthesis of their subjective experience (Kahneman and Klein, 2009)—rather than a more formal weighing of alternatives and probable outcomes in order to decide what to do and how to shape events as they unfold. Gary Klein, one of the founders of this movement, (2008) provides an excellent overview of the development of this approach to understanding “real-world” decision making.

Nobel laureate Daniel Kahneman and Klein (2009) describe the boundaries between situations when naturalistic decision making by experienced experts is necessary and appropriate and those in which a thoughtful analysis of probable outcomes is better. In a nutshell, NDM can be the only option in situations where one cannot assess critical aspects of the environment or the dynamic social system in flux. The caveat, however, is that NDM is best done by “true experts”, people “recognized in their professions as having the necessary skills and abilities to perform at the highest level (Shanteau, 1992: 255).”

Much of the research done thus far in naturalistic decision making has focused on teasing out how true experts make decisions in threat environments ranging from military combat information centers, wildfire command posts, nuclear power plants, and offshore oil platforms to neonatal intensive care units. In those studies, a technique called “Cognitive Task Analysis” (CTA) is employed. CTA uses intensive, days-long interviews of true experts by professionals trained in the technique. Interviewers tease out the tacit knowledge that the true experts use to experience a situation in flux, identify from memory a similar situation they’ve experienced and how they dealt with it successfully, and quickly do a mental simulation of whether that approach could be used—or modified for use in the current situation. If not, they rapidly reiterate the process until they come up with a tenable alternative to implement (see diagram at Klein, 2008:459). The results from CTA form the basis for incident analysis, policy and training development, and research in whichever endeavor is under study.

CTA can be used in policing (e.g., Lande and Klein, 2016), but it is requires highly trained research staff for interviewing and analysis, and is very time intensive for both the true experts being interviewed and those doing the interviewing. Although it can provide valuable insights that may be translated into policy or training, it doesn’t provide precise metrics for developing policy and training, evaluating performance, and refining practices. These are the tools
an organization needs in order to learn and adapt to change. Furthermore, policing agencies and the researchers who study them seldom have the kinds of resources needed for CTA-based problem analysis or development of policies and training. And police problem solving often must be done under great time pressure, for example, because of public outrages for change, the emergence of new threats, or budgetary crises.

The novel metric development process described here can provide results that are very similar to CTA by mining true experts’ implicit knowledge. It then applies well-validated scientific methods in a rapid, efficient, and cost-effective scaling process. The result provides objective, precise metrics for measuring what matters in a wide range of problem areas faced by contemporary policing organizations and the communities they serve.

Traditional Approaches to Measurement and Change

Traditionally, when an officer’s performance in an encounter was questioned or elicited an external complaint, police have assessed performance based on its 
outcomes, whether applicable laws and policies were followed, and the totality of the circumstances surrounding the encounter (e.g., see Graham v. Connor 490 U.S. 386 (1989)). These assessments were made by superior officers with substantial amounts of experience, sometimes with input from formal internal or external investigative reviews or panels of subject matter experts. If the officer’s performance was deemed substandard, consequences ranged from counseling, training, administrative or disciplinary measures to termination and prosecution. People who thought they had been wronged during encounters with police could institute civil proceedings, and political pressure could be exerted by elected officials, community activists, and the media. In more extreme cases, if probable cause was deemed to be present, judicial proceedings could be brought by local, state, or federal prosecutors that targeted the officer and/or the employer depending on the circumstances (Walker, 2001).

However, traditional responses and remedies have tended to ignore the dynamic and probabilistic nature of encounters between police and the public, focusing on outcomes, subjective eyewitness testimony and—more recently—videos that record an encounter from a limited perspective. This tendency to ignore the realities of police encounters can undermine justice itself (e.g., see Vila, 1992).

Justice and Human Performance

Our limited ability to realistically understand and measure the dynamic and probabilistic realities of police encounters with people has undermined both community justice and justice in the police workplace. As Princeton ethicist Kwame Anthony Appiah put it, “If you say somebody ought to do something, you must be supposing that it is something they can do” (2008: 22). It also has limited the scope of police performance research in two ways: (1) by encouraging a focus on macro-level assessments of organizational policies, practices, and training (e.g., Langworthy 1999; Davis et al., 2015); and (2) by limiting examinations of officers’ micro-level behaviors immediately before and during encounters to subjective, categorical and ordinal measurement. (See Mastrofski (2004: 109) for a commentary on the limited scope of our assessment of police performance.)

Police training also has been seriously undermined by our perceived inability to measure what matters in police encounters with the public—especially those that often result in the use of force. Every sworn police officer is trained to manage such encounters with the public, and certified as qualified by state standards. In order to be certified, they must demonstrate an ability to observe what occurs, weigh potential courses of actions against complex moral, legal, policy, and tactical considerations, decide, and act.

The problem is, no empirical connection has yet been established between what individual officers have been taught regarding how to manage such encounters and their ability to perform on the street. Instead, we rely on the outcomes of their encounters with the public—with a nod to “the totality of the circumstances”—or macro-level assessments of organizational performance. In the case of low base-rate phenomena such as officer involved shootings (Brown et al., 2001), objective systematic observation hasn’t been possible. At the individual level, post-hoc ethnographic research has been as close as we could get (e.g., Klinger, 2004; Pinizzotto et al., 2006, 2007), although systematic field observations of officer coercion has been done in specific cities (e.g., Terrill & Mastrofski, 2002).

Another dimension of this critical gap is the default assumption that there is a valid causal connection between the training we provide and the trainee’s future performance in an encounter with the public. The following quote from a
former president of the International Association of Directors of Law Enforcement Standards and Training illustrates this point:

Each [state peace officer standards and training agency (POST)] in cooperation with and supported by community leaders, elected officials, professional law enforcement administrators, academicians, and the directors' association, has established a standard that each officer has passed. These standards are not arbitrary, not based on lore, supposition, or wishful thinking; rather, each required characteristic has been identified and validated [emphasis added] as predictive of the officer's capacity to perform the job's essential functions. The ability to perform those same essential job functions serves as the basis for the officer's initial training. Career-long mastery of the evolving requisite skills is a POST requirement for certification. Each community can look upon its police officers with a sense of security and confidence, knowing that each officer has been certified by their POST agency. (Bradley, 2005)

Despite this authoritative assertion, officer training has not been shown to have a scientifically valid impact on individual officer performance in the field with regard to the types of police/public encounters that regularly lead to tragic outcomes and public outrage which, in turn, corrode public perceptions of the police. That relationship has yet to be tested in most cases—in part because of a lack of appropriate measurement tools. In Appiah’s terms, we have been supposing that the training we give officers enables them to manage encounters effectively without doing the science. Justice requires that we understand what is possible and assure that our training is valid.

Overview and Purpose

The goal of the research reported below was to create an efficient and relatively inexpensive process for developing timely, interval-level metrics that could be used to accurately measure what matters with regard to understanding the performance of individual police officers in encounters with the public, training for those encounters, and developing policy. The theoretical foundation for these metrics is Normal Accident Theory (Perrow, 1984) as adapted for use with police officers by Klinger (2005). The metrics development process we created was used in three different critical performance areas during a seven-year period.

After separately reviewing the relevant literature for each type of encounter, we developed metrics for measuring officer performance and the relative situational difficulty of encounters in three areas: Deadly force judgment and decision making (DFJDM, 2008-12); Tactical social interaction (TSI, 2012-13); and Crisis intervention teaming (CIT, 2013-14). In each case, we first used a form of concept mapping (Kane and Trochim, 2007) to derive important indicators of officer performance and situational difficulty from the implicit expertise of true experts in that performance area. The focus groups were selected and convened separately for each of the performance areas for which we developed metrics. After each, we used Thurstone scaling techniques (Thurstone and Chave, 1929; Thurstone, 1959) with diverse samples of expert trainers for that performance area to attach interval level values across each variable’s commonly encountered range of values. The general metric development process is described below.

METHODS

The two primary components of the metric development process were concept mapping and Thurstone scaling. For the purpose of clarity, we present each component’s research design, participants, procedures, and materials separately. Where appropriate, we include differences between the methods used when developing DFJDM, TSI, and CIT metrics in order to provide a sense of the flexibility of this technique.

Reverse Concept Mapping

Concept Mapping Research Design

Concept mapping is a widely used research method for identifying, visualizing and transforming key concepts into measurement scales in situations where no widely accepted, objective measurement criteria are available, but participants in the research are able to provide substantial amounts of subjective expertise about the issue of interest. As a first step, Vila was qualified as a concept mapping facilitator after participating in a three-day training course at
Concept Systems, Inc. in Ithaca, N.Y., conducted by William M. Trochim and Mary Kane, who pioneered this technique and developed the software we used to implement it (Kane and Trochim, 2007).

Although the concept mapping process often is used to extract latent concepts from the more concrete knowledge of subject matter experts, our goal was to do just the opposite—to obtain the measurable variables underlying the expert judges’ elaborate conceptualizations of police encounters with the public. We call this “reverse concept mapping” because we used the concept mapping process to dive under the abstract conceptualizations that produced so many differences of opinion among true experts (see Shanteau, 2001) to get at the measurable details that they agree are important. For example, in a deadly force encounter, many of these experts disagree about fundamentals such as how to define situational awareness, whether officers should endanger themselves in order to save a bystander, and whether officers tend to respond automatically in those encounters rather than go through decision making processes such as naturalistic decision making.

**Concept Mapping Participants**

We convened a diverse panel of true experts for each set of concept mapping focus groups. Although the ethics of research with human subjects\(^1\) preclude identification of the participants, the diverse expertise of participants in the true expert focus groups included:

**DFJDM**: 17 true experts with extensive experience in policing, firearms training and deadly force research, several with experience in all three categories; 15 of whom had extensive current or past experience in large, medium or small policing agencies in different parts of the United States; several of whom had worked in rural, suburban and urban settings; six who had come up through the ranks to become executives; one officer currently working patrol with slightly less than 15 years’ experience who was peer nominated as the “best street cop” in a mid-sized city; two women; and leaders from four major law enforcement training organizations.

**TSI**: 12 true experts from police, military, and academic backgrounds. Each participant had extensive experience in one or more of the following: cross-cultural operations in threat situations where the ability to win peoples’ support and de-escalate encounters was vital; studying naturalistic decision making; military special operations; combined military/policing operations; managing and interpreting communications with people whose world view can be extraordinarily different; and training curriculum development.

**CIT**: 18 diverse true experts drawn half from a medium-sized city’s criminal justice organizations (city police, county sheriffs, and corrections) and half from its community mental health professionals. Each had extensive experience in one or more of the following: responding to calls for service involving a mentally ill individual(s); police patrol tactics; police defensive tactics; police negotiation tactics; motivational interviewing; CIT curriculum development; treating mental illness and developmental disability; and emergency care for persons in crisis.

**Concept Mapping Procedures**

The concept mapping process followed well-defined steps across a two-day intensive focus group: the first day focused on encounter difficulty, and the second on officer performance. On the morning of the first day, participants came to a consensus about the most concise description of the goal of the encounter in question. These goals were the shorthand equivalent of the often-elaborate rules of engagement issued in police and military operations. For example, the DFJDM goal was “The goal of a police officer in a deadly force encounter is to accurately identify a threat and neutralize it while minimizing harm to bystanders, officers, and suspects.”

The focus group members then took turns nominating “difficulty” indicators or statements (variables that tend to make the likelihood of achieving the goal in question more challenging). After a statement was nominated, a facilitated group discussion followed until the statement was either modified to satisfy all of the participants or rejected by consensus. Once the group was satisfied that all of the critical dimensions affecting situational difficulty were addressed, overlapping statements were integrated based on additional discussion and duplicates were removed to create a final list of variable statements. Participants next sorted the difficulty statements into categories and rated them in terms of both their *importance* and the *frequency* with which they occur in operational settings. Based on the

\(^1\) These projects were approved by the authors’ Institutional Review Board.
sorting, maps, charts, and go-zones (bivariate graphs that identify statements that are above average for both variables, for example “importance” and “frequency”) were then created as visual aids for the focus group to use when determining which statements were most opportune (in terms of both being important and occurring frequently) with regard to achieving the goal in question.

On the second day, the process was repeated using the performance focus prompt to identify indicators or statements of officer behaviors that are likely to increase the likelihood of achieving the goal in question. The key difference between the days was that, once participants were satisfied that all of the critical dimensions affecting officer performance were addressed, performance statements were sorted into categories but rated on their importance and their feasibility (how difficult they are to do). As with difficulty statements, maps, charts, and go-zones were created as visual aids for the focus group to use when determining which statements were most opportune (in terms of importance and feasibility) with regard to achieving the goal in question.

THURSTONE SCALING

Scaling Research Design

The Thurstone equal-appearing interval scaling approach (see Miller & Salkind 2002; Thurstone & Chave 1929; Trochim 2006) is a method for assigning values to our statements by rating them on a Likert-type scale (e.g., 1=least impact on difficulty or performance, 7=greatest impact on difficulty or performance). This approach determines the meaning and magnitude of the different statements. Scores assigned to the statements during the Thurstone scaling process are calculated using Thurstone’s equal-appearing interval method with median weighting. For example, suppose true expert judges completed the performance prompt with this statement:

• “An element of a police officer’s performance in commonly encountered deadly force situations which increases the likelihood of achieving this goal [i.e., the DFJDM goal statement] is: maintaining a well-balanced stance.”

Then, the statement “maintaining a well-balanced stance” would be scored on a seven-point scale by hundreds of use-of-force instructors (see below). If the median value assigned by the use-of-force instructors to that statement was determined to be six, then six would be the value assigned to the overall performance score of an officer who maintained a well-balanced stance in a deadly force encounter.

Scaling Participants

The participants who were expert raters for the Thurstone scaling process across the development of the three different metrics were as follows:

DFJDM: Participants were 323 police officers from 209 different agencies across the United States who were use-of-force instructors. We used a snowball recruiting process, whereby experts from the concept mapping group recommended our survey to colleagues, police agencies and departments, and police unions and fraternal orders around the country.

TSI: Participants were 196 experienced law enforcement officers recruited from agencies across Washington State. Recruitment was done via their agencies’ training officers. Nearly all (90%) of the officers had more than five years of sworn police experience, their mean age was 44 years (SD 9.0), 89% were male, and 7% belonged to racial/ethnic minority groups. Three-quarters of them had received cultural awareness training, and 33% had prior military experience—nearly half with foreign deployment. (Note: Speed was essential for this DARPA-funded project.)

CIT: Participants were 499 police officers and mental health professionals from different agencies across the United States (and several law enforcement agencies from the UK and Canada). We used a snowball recruiting process, whereby SMEs from the focus group recommended our survey to colleagues, police agencies and departments, and fraternal orders around the country.

Scaling Procedures

Online surveys were prepared that contained the difficulty and performance statements with Likert-type scales of appropriate length (see discussion below in DFJDM results section). Scale length and range were specified to be as
consistent as possible with previous data collection and research on police encounters. The survey distribution process was designed to assure that participants’ identities were confidential (DFJDM) or anonymous (TSI, CIT).

Survey participants rated each statement for importance on a predetermined Likert-type scale. As was discussed previously, this process yielded interval-level scales for measuring the concepts judged to be most relevant for understanding both officer performance and the relative difficulty of different situations. For a full discussion of why this process yields interval-level data from ordinal-level expert rankings, see Thurstone and Chave (1929) or Trochim (2006).

RESULTS

Deadly Force Judgement and Decision (DFJDM) Making Metrics.

During the concept mapping workshop, the expert DFJDM participants generated 111 statements relating to the difficulty of a deadly force encounter, and 105 statements relating to officer performance in a deadly force encounter. Some of these statements needed to be broken into measurable items. For example, “distance from suspect” (relating to difficulty) was broken into: “the distance between the officer and the suspect is 0-5 feet,” “the distance between the officer and the suspect is 6-10 feet,” “the distance between the officer and the suspect is 11-20 feet,” “the distance between the officer and the suspect is 21-50 feet,” and “the distance between the officer and the suspect is greater than 51 feet.” These increments were chosen based on conventions in official statistics and previous research on officer-involved shootings. This resulted in 311 difficulty statements and 289 performance statements. The number of operationalized difficulty and performance statements was far greater than we had originally anticipated, thus smaller “sub” surveys were created for the DFJDM scaling process to avoid time-on-task effects. During the Thurstone scoring process, the scores assigned to difficulty statements by the use-of-force instructors ranged from 1 (no impact on difficulty) to 7 (highest impact on difficulty). Consistent with the Thurstone scaling process, the median value given to each statement by the raters then was assigned as that statement’s value.

Examples of statements that received the highest difficulty score (7) are:

- “The portion of the suspect exposed to the officer’s view is 0%”
- “There are eleven or more hostages under direct threat at the scene”
- “There are eleven or more suspects present at the scene”

Examples of statements that received the lowest difficulty score (1) are:

- “The suspect’s apparent physical strength is weak”
- “The suspect has no visible gang identifiers”
- “Potential threats can come from only one direction”

The scores assigned to performance statements by the use-of-force instructors ranged from -6 (extremely negative impact on performance) to 0 (no impact on performance) and up to +6 (extremely positive impact on performance).

Examples of statements that received the highest performance score (+6) are:

- “The officer used proper tactical movement”
- “The officer makes timely decisions regarding pre-assault indicators”
- “The officer accurately identified multiple opponents”

Examples of statements that received the lowest performance score (-6) are:

- “The officer has averaged 0-3 hours sleep in recent days”
- “The officer doesn’t expect to be involved in a deadly force situation”
- “The officer doesn’t apply deadly force rules of engagement (laws and policies) in a combat situation”

2 Our determinations were guided by advice from three of the true experts initially selected for the focus group who were unable to attend at the last minute.
Tactical Social Interaction (TSI) Metrics

During the Concept Mapping process, the expert TSI participants generated 147 difficulty indicators and 78 performance indicators. During the Thurstone scoring process, the scores assigned to difficulty statements by the TSI survey respondents ranged along a seven-point scale from “Makes It Easy” to “No Effect” to “Makes it Hard”. Consistent with the Thurstone scaling process, the median value given to each statement by the raters then was assigned as that statement’s value.

Examples of statements that received the highest difficulty score (Makes it Hard) in tactical social interactions are:
- “General animosity toward law enforcement”
- “The civilian using deception/manipulation”
- “The civilian resenting the officer's presence in the area”

Examples of statements that received the lowest difficulty score (Makes it Easy) are:
- “One civilian is involved in the encounter”
- “The civilian’s facial expression is (smiling)”
- “The civilian is an older adult”

The scores assigned to performance statements by the TSI survey respondents ranged along a seven-point scale from “Not Important at All” to “Extremely Important”.

Examples of statements that received the highest performance score (7) are:
- “Understanding rules of engagement”
- “Maintaining emotional control”
- “Maintaining a position of tactical advantage”

Examples of TSI performance statements that received a lower score (4) are:
- “Gesturing toward the civilian over aggressively”
- “Mirroring what civilians do in greeting”
- “Establishing common ground”

Crisis Intervention Team (CIT) Metrics

During the Concept Mapping process, the expert CIT participants generated 90 difficulty indicators and 112 performance indicators. During the Thurstone scoring process, the scores assigned to difficulty statements by the CIT survey respondents ranged from 3 to 6 even though the scales also presented response categories 1 and 2. Consistent with the Thurstone scaling process, the median value given to each statement by the raters then was assigned as that statement’s value.

Examples CIT of statements that received the highest difficulty score (6) are:
- “Inadequate staffing being available to respond to the crisis”
- “Not knowing where the person in crisis is located”
- “Having the opportunity for multiple persons to go into crisis (e.g., a bomb explosion, school shooting)”

Examples of CIT statements that received the lowest difficulty score assigned (3) are:
- “The presence of law enforcement officers escalating the crisis”
- “The person in crisis having immediate medical needs (e.g., seizures)”
- “The person in crisis appearing to be physically imposing (e.g., very muscular)”

The scores assigned to performance statements by the CIT survey respondents ranged from 1 to 6.

Examples of CIT statements that received the highest performance score (6) are:
- “Seeking accurate information about the situation before arrival”
- “Reading non-verbal cues of the person in crisis”
- “Adapting based on changing level of threat”

Examples of CIT statements that received the lowest performance scores (1 or 2) are:
- “Knowing key mental health terms (e.g., legal acronyms)”
- “Mentally rehearsing before the encounter”
- “Adapting response based on [the officer’s] own current emotional status”
Practical Application of the Metrics

The metrics are straightforward to use for scoring scenario difficulty and officer performance. Our experience thus far has been that, with less than a day of training to learn the meaning of policing terms used in the metrics and test to criterion levels of rating reliability, scorers can use the metrics fairly efficiently.

Difficulty Scoring

Difficulty variables for each encounter scenario measure the extent to which the situational difficulty of a scenario reduces the likelihood that an officer will be able to achieve the goal of an encounter (see Concept Mapping Procedures subheading above). They are scored on a seven-point, positive scale ranging from 1 (no impact on difficulty) to 7 (highest impact on difficulty). So the difficulty value for a particular scenario is a fixed, positive number. The scenario difficulty value is calculated by: (1) identifying which of the possible difficulty variables were present in it, (2) using the difficulty metrics to assess the magnitude of each variable present to calculate its difficulty score, and then (3) summing the scores for all present variables.

Performance Scoring

Performance variables in each encounter scenario measure the extent to which an officer’s behavior in that encounter affects the likelihood that he or she will be able to achieve the encounter’s goal (see Concept Mapping Procedures subheading above). They are scored on a 13-point scale ranging from -6 (extremely negative impact on performance) to +6 (extremely positive impact on performance). So each of the types of action an officer does during an encounter that is one of the behavioral variables is scored separately. For example, if an officer responds in a timely enough fashion to pre-assault indicators, he or she might receive a score as high as +6. But if the assault began before he or she noticed the pre-assault indicators, the score might be as low as -6. The raw performance score for the encounter would then be calculated as the sum of these positive and negative scores. Then, the final performance score can be calculated by using the raw score as the numerator and the maximum possible performance score (+6 \cdot (\sum \text{scored behaviors}) as the denominator. Thus, performance scores are expressed as a proportion of perfect performance in all measured behaviors.

It is important to stress that we do not expect that a perfect performance score is possible except, perhaps, in the least difficult encounters. Naturalistic decision making involves the use of successive expert approximations of the nature of a dynamic, complex, and low-information situation as it unfolds. It relies heavily on expertise, behavioral flexibility and ability to improvise. So, “good enough, quick enough” tends to be what is critical in order to achieve a desirable, or at least acceptable outcome. Our goal is to provide a relatively precise and reliable measure that can be used to gain a better understanding of police performance in encounters with the public. Metrics for assessing DFJDM, TSI, and CIT encounters may be obtained from the authors on request.

Example of DFJDM Metrics in Use

Table 1: Example of scoring process.

<table>
<thead>
<tr>
<th>All Items</th>
<th>Weighted Score (A)</th>
<th>Possible to Achieve (B)</th>
<th>Potential Score (C = A x B)</th>
<th>Behavior Observed (D)</th>
<th>Actual Score (E = C x D)</th>
<th>% Performance (Sum E / Sum C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 2</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>Yes</td>
</tr>
<tr>
<td>Item 3</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>N/A</td>
<td></td>
<td>Not Achievable</td>
</tr>
<tr>
<td>Item 98</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>N/A</td>
<td></td>
<td>Not Observable</td>
</tr>
<tr>
<td>Item 99</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Item 100</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Max Score (Sum C) 22
Actual Score (Sum E) 15
68%

In a recent study, conducted by Vila et al. (2015), police officers (n=80) completed 2,082 custom high definition DFJDM video scenarios simulations. Errors in the simulator, like those in real life, were rare. The officers made 103 errors (sum of all errors: shooting unarmed, not shooting armed, shooting before threat was apparent, and hitting
innocent bystander) which accounted for 4.9% of all scenarios completed. To maximize the training potential of DFJDM simulators a finer grain look at officer performance is necessary. Performance is measured by assessing all the actions officers could have taken in a given scenario to ascertain a maximum possible score. Then the officer’s observable behavior is weighted (as per the Thurstone scaling process) and scored. The observed score is then divided by the maximum possible score to obtain the % performance score. Table 1 shows the scoring process.

Using this process, the mean performance score was 71% (SD=.18). As can be seen in figure 1 below, the performance of officers across 9 different scenario types varies greatly. Applying the metrics to officer performance can therefore inform training needs and identify skill gaps at the type-of-encounter level. Drilling down further can reveal specific behaviors that are desired, but frequently not displayed at the group, or even individual, level.

**Figure 1: DFJDM metrics applied to 2,082 video simulations completed by police officers (n=80).**

**DISCUSSION**

**Summary**

The novel metric development process described here provides a practical method for rapidly creating tools for assessing individual officer performance in different types of encounters with the public. With regard to the speed with which this metric development process can be applied, we note that the time required to develop and employ each of these metrics declined rapidly as we gained experience; from three years (DFJDM), to six months (TSI), and then to six weeks (CIT). The CIT metrics and a week-long, multi-agency training program were developed in response to community outrage, and a subsequent independent use of force commission recommendation, over an in-custody death caused by the way officers handled a developmentally challenged person while responding to a robbery-just-occurred call. As was described previously, the metrics also are efficient to employ once scorers become familiar with the scoring items.

The metrics developed using these techniques provide a solid initial platform from which policies can be prototyped in response to public pressure, curricula can be established for training officers to implement that policy, assessing their performance pre-, peri-, and post-training, and later in the field. The knowledge gained could then be used to evaluate the effectiveness of policy, practice and training, and refine any or all three until they appear to be satisfactory. In a learning organization, this process likely would continue to evolve in order to adapt policies and practices in a changing world.
This highlights another difference between our metrics and more traditional instruments, inventories, and assays. Our approach is not comprehensive—it does not provide measures for all relevant variables in all types of scenarios. And its generalizability is limited by the diversity of the true-expert focus group and the representativeness of the population that rated the magnitude of each set of scales’ increments. However, the metrics produced using this approach are good enough to provide a starting place for measuring what matters when it must be done now. Our process provides a “good-enough” starting point for what is necessarily an evolutionary process. Like biological evolution, these sorts of approaches to innovation are very good at finding optima in complex, dynamic systems, but they must have somewhere plausible to begin. The metrics developed thus far can be refined through repeated iterations of the process with potential offshoots, or by creating entirely new metrics to measure similar problems that arise.

The metrics we have developed with this process are unlikely to capture all of the relevant variables for understanding police performance in everyday encounters with the public. But they do provide a more plausible, extensive, and precise set of measures than was ever before available, and that set can be refined and expanded over time. These metrics lay a rough empirical foundation for the study of individual officer performance in encounters with the public. That is the critical juncture where state and citizen interact, services are provided, and justice flows or fails. We can no longer say “That’s too complex to measure.” At a minimum, we can finally begin to assess which of the things that officers do increase the likelihood of a good outcome in encounters between police and public and which reduce it.

Conclusions

This approach to developing metrics can have both empirical and political advantages. At times when rhetoric and outrage tend to obscure evidence and logic—and impatience encourages ineffective expedients—the metrics process can be used to help focus attention on causal processes rather than just tragic outcomes. In effect, it could be used as an impartial, science-based, technique for quickly facilitating conflict mediation. Instead of focusing on assigning blame and punishment for a particular event’s outcome, which necessarily tends to require slow-moving adjudication, a diverse focus group of true experts could identify the difficulty and performance variables in such an encounter, then suggest the most important issues to focus upon. This puts the focus on evidence-based problem solving, identifying the most important variables associated with a problematic type of encounter, and considering causal processes that provide traction for solution, rather than who-did-what. This would help focus the public discourse on solutions, relieving some of the pressure on public officials and enabling policy development and measurement specification to proceed using concrete, unambiguous terms. This is a naturalistic process that focuses on a research question “What do we know now, and what can be done about it here and now?” It acknowledges that policies and training and interventions are hypotheses which must be tested and refined. Often, real world crises of the sort discussed here can’t wait for gold standard randomized control trials, and they need not wait.

REFERENCES

Shanteau, J. (2001). What does it mean when experts disagree? (In Salas, E. & Klein, G. (Eds.), Linking expertise and naturalistic decision making (pp. 229-244) Mahwah, N.J.: Lawrence Erlbaum Associates.)
Thurstone, L. L. & E. J. Chave. (1929) The measurement of attitude: A psychophysical method and some experiments with a scale for measuring attitude toward the church. (Chicago: University of Chicago Press.)

Legal Cases