How police officers perform in encounters with the public
Measuring what matters at the individual level

Bryan Vila
Department of Criminal Justice and Criminology, Washington State University, Spokane, Washington, USA
Stephen James
Sleep & Performance Research Center, Washington State University, Spokane, Washington, USA, and
Lois James
College of Nursing, Washington State University, Spokane, Washington, USA

Abstract
Purpose – The purpose of this paper is to develop and describe the implementation of a novel method for creating interval-level metrics for objectively assessing police officer behaviors during an encounter with the public. These behaviors constitute officer performance and affect the probability of desirable encounter outcomes. The metrics measure concrete, micro-level performance in the common types of complex, dynamic, and low-information police-public encounters that often require immediate action using “naturalistic” decision making. Difficulty metrics also were developed to control for situational variability. The utility of measuring what officers do vs probabilistic outcomes is explored with regard to informing policymaking, field practice, and training.

Design/methodology/approach – Metric sets were developed separately for three types of police-public encounters: deadly force judgment and decision making, cross-cultural tactical social interaction, and crisis intervention. In each, “reverse concept mapping” was used with a different diverse focus group of “true experts” to authoritatively deconstruct implicit concepts and derive important variables. Variables then were scaled with Thurstone’s method using 198 diverse expert trainers to create interval-level metrics for performance and situational difficulty. Metric utility was explored during two experimental laboratory studies and in response to a problematic police encounter.

Findings – Objective, interval-level metric sets were developed for measuring micro-level police performance and encounter difficulty. Validation and further refinement are required.

Research limitations/implications – This novel method provides a practical way to rapidly develop metrics that measure micro-level performance during police-public encounters much more precisely than was previously possible.

Originality/value – The metrics developed provide a foundation for measuring officers’ performance as they exercise discretion, engage people, and affect perceptions of police legitimacy.

Keywords Procedural justice, Performance measures, Evaluation metrics, Naturalistic decision making, Police discretion and accountability, Police policy and training

Paper type Research paper

Introduction
Some of the most divisive and corrosive issues in contemporary American society are anchored in police interactions with the public. If the past few years have taught us anything, it is that the most important aspects of police performance are how individual officers behave on the streets, and how the public perceives those actions. Unfortunately, our ability to objectively measure how individual officers perform as they manage encounters with the public is woefully inadequate. This is a critical problem, as Mastrofski (2002) 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” (p. 113). The research reported here developed a novel method for measuring the micro-level details of police behavior. The resulting interval-level metrics make it
possible to readily and objectively assess the dynamic sequence of officer behaviors during an interaction with the public that constitute performance and affect the probability of a desirable outcome[1].

Such metrics are important because most police performance measures on the streets are post hoc; they focus on outcomes, rather than what police do or do not do to achieve them. This is a coarse approach to a nuanced problem because police encounters with the public are complex social interactions and their outcomes are inherently probabilistic. 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. This lack of precision undermines police training, evaluation research, and policymaking.

Similarly, our limited ability to understand and measure the dynamic and probabilistic realities of police encounters tends to undermine legitimacy on both sides of the police-citizen divide. Cops know well what the public regularly forget: sometimes you do everything right and it all goes wrong; other times you do it all wrong, and get a great outcome. Luck – probability – matters on the street, and that fact is difficult to factor into how we hold police accountable for exercising discretion. As Princeton ethicist Kwame Anthony Appiah (2008) put it, “If you say somebody ought to do something, you must be supposing that it is something they can do” (p. 22). Our suppositions about what police can do in encounters with the public often ignore the realities of social encounters.

Dynamic social encounters
How an officer behaves toward the people he or she encounters generates a cascade of responses, counter responses and interactions with participants and bystanders (e.g. see Goffman, 1969). This dynamic network of interactions evolves rapidly, surging and waning as each action spawns others, reinforcing some possible paths for response and ignoring 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 employ in order to influence others in the encounter and guide its course toward a desired outcome (Bakeman and Gottman, 1986; Mastrofski et al., 2015).

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 that evaluations of officer performance must consider the situational difficulty of the environment in which encounters between police and the public unfold.

In addition to environmental challenges, the dynamic forces at play in social encounters also must be considered. As encounters become more situationally and socially interconnected, or as the pace of interactions accelerates, they become increasingly difficult to understand, control, and predict (see Klinger, 2004, 2005). This increase in turbulence (McCann and Selsky, 1984) raises the probability of an unforeseen or unforeseeable catastrophe (Perrow, 1984). Catastrophic outcomes in social encounters at the micro level become humanly 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 is going to happen next and how best to respond. More detailed explanations of these dynamics may be found in Eubank and Farmer (1990), Holland (1992, pp. 184-185), Mitchell (2009), Vila (2010), and Eagleman (2011).
The real-world implication of these dynamics is that encounters between the police and the public are probabilistic. Ideally then, in order to assess the propriety and overall justness of officers’ actions, we need to assess what they do in an encounter – the relative difficulty of the situation and environment, and the limits of human performance – not just how things turn out.

**Naturalistic decision making (NDM)**

Police officers often rely on “expertise” 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 in the sorts of complex, dynamic, fast-paced, and low-information situations that often characterize their encounters with the public. By expertise, we mean an officer’s skilled intuition based on a synthesis of his or her subjective experience (Kahneman and Klein, 2009). NDM, which was developed during the late 1980s, is one of the most promising approaches for understanding how people make decisions in these kinds of situations[2].

In a high-profile paper on decision making in situations requiring intuition and expertise, Nobel laureate Kahneman and Klein (2009) (one of the founders of the NDM movement) debated the boundaries between situations when NDM by high-level experts is necessary and appropriate, and those in which a thoughtful analysis of probable outcomes is better. They concluded that NDM can be the only option in situations where one cannot assess critical aspects of the environment or the dynamic social system in flux. Their caveat, however, is that NDM is best done by “true experts,” people who are “recognized in their professions as having the necessary skills and abilities to perform at the highest level” (Shanteau, 1992, p. 255).

Much of the research done thus far in NDM has focused on understanding 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 have 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 this process until they come up with a tenable alternative (see diagram at Klein, 2008, p. 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, and may provide valuable insights that can be translated into policy or training (e.g. Lande and Klein, 2016). However, it is expensive and time-intensive for both the true experts being interviewed and those doing the interviewing and analysis. This can be impractical, because police leaders often must solve problems under great time pressure due to public outcries for change, the emergence of new threats, or budgetary crises. CTA also does not 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.

The novel method for metric development described here can quickly provide results that are very similar to CTA by mining true experts’ tacit knowledge. But it also applies well-validated scientific techniques in a rapid, efficient, and cost-effective scaling process. The result provides timely, objective, and precise metrics for measuring what matters in a wide range of problem areas faced by contemporary policing organizations and the communities they serve. For example, the three sets of metrics reported here can arguably be used to help address each of the six “pillars”[3] identified by the President’s Task Force on 21st Century Policing (2015) for improving the ways that policing agencies “interact with
and bring positive change to their communities” (p. 1; see also Office of Community Oriented Policing Services, 2016; Tyler et al., 2015). The core of those pillars is how officers behave during interactions with the public.

Traditional approaches to measurement and change

Traditionally, when an officer’s performance in an encounter was questioned or elicited an external complaint, the justice system has assessed performance based on three things: the encounter’s outcome, whether applicable laws and policies were followed, and the totality of the circumstances surrounding the encounter (e.g. see Graham v. Connor 490 US 386 (1989)). However, traditional responses and remedies often tend to ignore the dynamic and probabilistic nature of encounters between police and the public. Instead, they often focus on subjective eyewitness testimony based on recall 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).

Our limited ability to realistically understand and measure the dynamic and probabilistic realities of police encounters with people has limited the scope of police performance research in two ways: by encouraging a focus on macro-level assessments of organizational policies, practices, and training (e.g. Langworthy, 1999; Davis et al., 2015); and by limiting examinations of officers’ micro-level behaviors immediately before and during encounters to subjective, categorical, and ordinal measurement. The US National Research Council’s 2004 review of police policy and practices in encounters made it clear that police performance measures tend to be ill-defined, often focus on outcomes, and have weak or unknown relationships to police practices or behavior (Skogan and Frydl, 2004).

Police training 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 first must observe an incident, and then demonstrate an ability to weigh potential courses of action against complex moral, legal, policy, and tactical considerations, make a decision, and act. As the NRC review found, 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 (Skogan and Frydl, 2004, pp. 109-154).

Perhaps more disturbing is the pervasive assumption among practitioners and the public that there is a valid causal connection between the training and performance in the field. As a former president of the International Association of Directors of Law Enforcement Standards and Training asserted in The Police Chief a year after the NRC report was issued:

Each [state peace officer standards and training agency] 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. Careerlong mastery of the evolving requisite skills is [...] [required] for certification (Bradley, 2005).

Bradley’s assertions reveal the blind spot created by our inability to measure police performance in encounters with the public at the appropriate scale. We cannot measure what matters, so we measure what we can and assume that is good enough. We rely on the outcomes of those encounters, with a nod to “the totality of the circumstances,” or to assessments of organizational performance based on aggregate behavioral and situational variability. This gap in our knowledge has been especially dangerous in the case of low
base-rate phenomena such as officer-involved shootings (Brown et al., 2001) where objective, systematic observation at the individual level has not even been possible. Generally, post hoc ethnographic research has been as close as we could get (e.g. Klinger, 2004; Pinizzotto et al., 2006, 2007), although some systematic field observations of officer coercion have been done in specific cities (e.g. Terrill and Mastrofski, 2002).

The current study
The goal of the research reported below was to create an efficient and relatively inexpensive method for developing timely, interval-level police performance metrics for measuring challenging police-public encounters that often require NDM. This is a formidable challenge because the settings and situations in which police-public encounters occur are extremely variable and difficult to predict. Encounters’ complex, self-reinforcing, and often tightly coupled social dynamics are probabilistic. Each encounter is unique at the micro level because the characteristics of both the people and the setting can differ so widely. This made a brute force approach to metric development impractical. However, intuitive expertise does exist. As novice police officers quickly realize during field training, experienced officers tend to be much more adept at managing encounters in ways that lead to good outcomes than rookies. And the best street cops can be amazingly good at their jobs, even though things sometimes go bad for them, too.

Our approach was to follow the well-established process for quantitative rationalization of subjective measures into interval- or ratio-level metrics that Thurstone and other psychometric and sociometric pioneers began in the 1920s (Thurstone, 1959, pp. 3-11 and 232). However, we sought to dramatically streamline the construction of a large suite of scales that measured what officers do in encounters. Scales were constructed from true experts about police performance in encounters with the public using a newer approach, computer-aided concept mapping. Then traditional Thurstone scaling methods were used to have hundreds of experienced police trainers assign values to the increments in the scales in a manner that produces interval-level measures.

Similarly constructed difficulty metrics were developed to control for environmental and situational factors. These metrics were intended to make it possible to accurately measure what matters at the micro level for understanding and assessing the concrete behaviors of individual police officers during encounters with the public, training for those encounters, developing policy, and conducting research. Although the purpose of this paper is to report on the metric development process, illustrative examples are provided for the three sets of metrics that we have developed thus far: deadly force judgment and decision making (DFJDM, 2008-2012); tactical social interaction (TSI, 2012-2013); and crisis intervention teaming (CIT, 2013-2014)[4]. The general metric development method is described below.

Methods
The novel metric development method is built with two primary components, concept mapping (reversed) 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 techniques used when developing DFJDM, TSI, and CIT metrics in order to provide a sense of the flexibility of the method. Figure 1 provides an overview of this process.

Reverse concept mapping
Concept mapping research design. Concept mapping is a widely used method for identifying and visualizing key concepts associated with a topic of interest, and sometimes for transforming them into measurement criteria. It is commonly used in situations where no
widely accepted and objective measurement criteria are available, but participants in the process are able to provide substantial amounts of subjective expertise. As a first step, a member of our team (BV) was qualified as a concept mapping facilitator after participating in a three-day training course at Concept Systems, Inc. in Ithaca, NY. The course was conducted by William M. Trochim and Mary Kane, who pioneered this use of the technique and developed the software we used to implement it (Kane and Trochim, 2007; see also Novak and Gowin, 1984).

As the name suggests, concept mapping often is used to extract latent concepts from the more concrete knowledge of a diverse group of subject matter experts. However, our goal was to do just the opposite. So we “reversed” concept mapping to derive concrete, measurable variables from the abstract concepts used by true expert judges to understand police encounters with the public. Reverse concept mapping helped the focus group deconstruct the abstract conceptualizations that produced so many differences of opinion among them[5] and made it possible to obtain consensus about important measurable behaviors and external factors. For example, many true experts about police deadly encounters 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 intentional cognitive decision-making processes such as NDM.

Focus group management. The success of our research hinged on effectively managing highly diverse groups of true experts – many of whom were on different sides of long-running debates about police encounters, tactics, training, practices, and policies. Uber expert officers, researchers and others with the deep insight and experience required for NDM in police work tend to be very strong personalities. The challenge was to keep them in a closed setting for two full days of focused, intense discussion. This required a firm yet impartial facilitator who also was expert on the issues, but had to be kept from biasing the group (e.g. by inadvertently asking a leading question in an attempt to gain

---

**Figure 1.**
A novel method for developing interval-level metrics for measuring performance requiring naturalistic decision making in high-risk situations as well as the relative difficulty of each situation.
clarity on a participant’s statement). We addressed this problem by having core members of the research team members serve as referees and “throw a flag” on the facilitator to minimize research bias. The gleeful arm waving that accompanied these interruptions lightened the intensity of group interactions and helped make it easier for participants to accept the facilitator’s nudges to move things along or broker compromise.

**Concept mapping participants.** We convened a diverse, separate, and unique panel of true experts for each of the three sets of concept mapping focus groups. Although the ethics of research with human subjects[6] preclude their identification, the participants’ diverse and overlapping expertise included:

- **DFJDM (2008)** focus group included 17 true experts with extensive experience in policing, firearms training, or deadly force research – several of whom had substantial experience in all three categories. Their occupational ranks/roles ranged from executive to a peer-nominated “best street cop.” Among the four researchers in the group, three had worked as sworn police officers.
- **TSI (2012)** focus group included 12 true experts with extensive experience in police, military, or relevant academic fields[7]. Each 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 NDM; military special operations; combined military/policing operations; and managing and interpreting communications with people whose world view can be extraordinarily different from one’s own.
- **CIT (2013)** focus group included 18 diverse true experts, including nine from a medium-sized city’s criminal justice organizations (city police, county sheriffs’ deputies, and corrections officers), and nine from its community mental health professionals. Each of these participants had extensive experience in one or more of the following: responding to calls for service involving a mentally ill individual (or individuals); police patrol tactics; police defensive tactics; police negotiation tactics; motivational interviewing; CIT curriculum development; treating mental illness and developmental disability; and emergency care for people in crisis.

**Concept mapping procedures.** The reverse concept mapping process followed well-defined steps across a two-day intensive focus group. Day 1 focused on encounter difficulty, and the following day 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. As was to be expected given the diversity of the experts, this process took approximately two hours for each focus group, but was critical to the metric development process. These goals were the shorthand equivalent of the often lengthy and elaborate rules of engagement issued in police and military operations. For example, the goal statement from the DFJDM focus group 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.”

Focus prompts based on the goal statements were then developed for guiding the difficulty and performance indicator discussions. For example, for the DFJDM goal:

- The difficulty focus prompt was: “An element of deadly force situations commonly encountered by police officers that increases the difficulty of achieving this goal is […]”
- The performance focus prompt was: “An element of a police officer’s performance in commonly encountered deadly force situations that influences the likelihood of achieving this goal is […]” The slides used in our focus group sessions also included performance statement examples that helped keep participants focused on the need for police officers to accurately identify a threat and neutralize it while minimizing harm to bystanders, officers, and suspects.”
to identify concrete, measurable indicators rather than the complex concepts with which they usually thought about performance in the situations being specified (e.g., situational awareness, command presence, etc.).

The focus group members then took turns nominating “difficulty” indicator statements for 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 statements.

Next, working on their own, participants sorted the difficulty statements into self-defined categories (e.g., suspect characteristics, suspect behaviors, and environmental factors). Each of them then rated each statement in terms of both its importance and the frequency with which it tends to occur in operational settings that commonly lead to officer-involved shootings. The research team used the participants’ sorting data to create maps, charts, and go/no-go zone charts (quadrant graphs that identify statements that are above average for both importance and frequency). These were used as visual aids for the focus group as they discussed which statements were the most important for achieving the goal in question. This also helped assure that the diverse concepts each participant had been concerned with during the day’s focus group were covered by the final statement sets.

Officer performance was assessed on the second day using the same concept mapping processes, but using the performance focus prompt instead of the prompt for difficulty. In this instance, the objective was to identify indicators or statements for officer behaviors that could increase the likelihood of achieving the goal in question.

*Concept mapping materials.* Specification and real-time prioritization and assessment of statements for each variable during the focus group was done using CS Global-Gold software (Concept Systems, Inc., Ithaca, NY, www.conceptsystems.com). This software allowed the true experts to independently and simultaneously sort and rate the statements on computers provided during the focus group. Once the focus group was completed, each variable was divided into Likert-type increments.

The concept-mapping process was used to fulfill all of the steps in the scaling method described by Thurstone (1959, p. 232) with the exception of data collection and analysis associated with calculating the value of statement indicators.

*Thurstone scaling*

The activities described in this section were used to fulfill the remaining scaling steps described by Thurstone (1959, p. 232).

*Scaling research design.* Thurstone’s equal-appearing interval scaling approach was used with median weighting to calculate values for the increments of each difficulty or performance indicator variable (see Miller and Salkind, 2002; Thurstone and Chave, 1929; Trochim, 2006). These values made it possible to estimate the extent to which each indicator variable influenced the probability of a desirable outcome. This well-validated method is based on the idea that one can develop objective, interval-level scales about indicator variables by obtaining subjective values from a sufficiently large and representative population of knowledgeable experts[8].

*Scaling participants.* The participants for the Thurstone scaling part of the metric development processes for the three different metrics were as follows:

- DFJDM: 323 police officers from 209 different agencies across the USA who were use-of-force instructors.
• TSI: 196 experienced law enforcement officers recruited from agencies across the state of Washington. Three-quarters of them had received cultural awareness training, and 33 percent had prior military experience – nearly half with foreign deployment.

• CIT: participants were 499 police officers and mental health professionals from different agencies across the USA.

Scaling materials. We used Survey Monkey (www.surveymonkey.com), an online survey service, to enable efficient scoring of variable statements by participants. After all rating was completed, Survey Monkey generated an output that was uploaded into Excel and SPSS. Survey Monkey was also used to gather demographic information from each participant.

Scaling procedures. We used a snowball recruiting process, whereby experts from each concept mapping group recommended (via e-mail) the survey from their group to colleagues, agencies and departments, as well as unions and fraternal orders around the USA. We also posted links to the surveys on various department and agency websites.

When participants accessed the surveys, they were instructed to rate each indicator statement on a Likert-type scale (e.g. 1 = least impact on difficulty or performance, 7 = greatest impact on difficulty or performance). The magnitude of each statement was then estimated using the median value assigned by the population of experts.

For example, survey participants might rate the following DFJDM performance indicator: “An element of a police officer’s performance in commonly encountered deadly force situations that increases the likelihood of achieving this goal is maintaining a well-balanced stance” with a value of 6 on a seven-point scale. Then the variable specified in that statement would be scored on a seven-point scale by hundreds of use-of-force instructors (see below). The median value assigned by hundreds of survey participants then becomes the weighted value for maintaining a well-balanced stance in a deadly force encounter.

Results
Metrics for assessing DFJDM (Vila et al., 2012), TSI (Vila et al., 2014), and CIT (James and James, 2017) encounters may be obtained from the authors on request. Although the purpose of this manuscript is to present the method for developing these types of metrics, some illustrative examples are provided below.

Note that the Likert scales differed between the performance metrics (DFJDM, TSI, and CIT). These differences were based on expert consensus about how much variation one could expect in performance[9]. For example, the items proposed during the DFJDM performance focus group included positive and negative items that had life-or-death consequences, thus we selected a broad range (−6 to +6). The items proposed during the TSI performance focus groups, however, were all positive and ranged from no impact to large impact on performance, so we selected a positive range (1-7). The items proposed during the CIT performance focus group incorporated both positive and negative items, but we determined that a narrower range was appropriate (−4 to +4) as they did not vary to the same extent as the DFJDM items.

Focus group scales
DFJDM metrics. During the DFJDM concept mapping focus group, the true experts generated 111 statements relating to the difficulty of a deadly force encounter, and 105 statements relating to officer performance in a deadly force encounter. Examples included “the suspect has visible gang identifiers” (difficulty) and “the officer accurately identifies multiple suspects” (performance). During the Thurstone scoring process, the scores assigned to difficulty statements by the use-of-force instructor raters ranged from 1 (no impact on difficulty)
to 7 (highest impact on difficulty). The scores assigned to performance statements ranged from $-6$ (extremely negative impact on performance) to 0 (no impact on performance), and up to $+6$ (extremely positive impact on performance). Consistent with the Thurstone scaling process, the median value given to each statement by the raters then was assigned as that statement’s value.

**TSI metrics**

During the TSI concept mapping focus group, the true experts generated 147 difficulty indicators and 78 performance indicators. Examples included “the civilian is visibly hostile” (difficulty) and “the officer explains the purpose of the encounter” (performance). During Thurstone scaling the expert TSI survey respondents were given a seven-point scale for difficulty items ranging from 1 (no impact on difficulty) to 7 (highest impact on difficulty). The same seven-point scale was used for performance ranging from 1 (no impact on performance) to 7 (extremely positive impact on performance).

**CIT metrics**

During the CIT concept mapping focus group, the true experts generated 90 difficulty indicators and 112 performance indicators. Examples included: “the person in crisis cannot communicate due to language barriers” (difficulty) and “demonstrating concern for the person in crisis’ safety” (performance). Difficulty scales scored by the expert raters ranged from 1 (no impact on difficulty) to 7 (highest impact on difficulty). Performance scales ranged nine points, from $-4$ (strong negative impact on performance) to 0 (no impact on performance), and up to $+4$ (strong positive impact on performance).

**Using the metrics**

The metrics are straightforward to use for scoring encounter difficulty and officer performance. Our experience thus far has been that scorers can be trained to use the metrics fairly efficiently within less than a day. During the training they learn the meaning of policing terms used in the metrics, then practice scoring for several hours before achieving criterion levels of $\leq 5$ percent errors for both within-rater and inter-rater reliability. Any indicator statements that are not completely objective (e.g. “the officer established common ground”) are accompanied by examples such as “the officer asked the civilian questions to see if they had anything in common, and used that to make a connection with the civilian – so that the civilian saw the officer as similar to him or her”.

**Scoring of difficulty**

Difficulty metrics are applied by: reviewing all information one has about the encounter (e.g. incident reports, camera footage, etc.), identifying which of the possible difficulty variables were present, referring to the difficulty metrics to assign a difficulty score for each observed variable, and summing the scores for all present variables to calculate the encounter’s relative difficulty.

**Scoring of performance**

The performance metrics are applied by first identifying which performance indicators are possible for an officer to achieve. This step is important to ensure that an officer’s performance score is not hampered by something he or she could not have done. For example, if officer performance in a training simulator is being scored, and the simulator does not provide the officer with cover, then the indicator “the officer made use of cover” would not be selected. Next, relevant performance indicators are selected and all relevant information one has about the officer’s performance is reviewed (e.g. incident reports,
camera footage, training simulator video, etc.). Then, the performance metrics are used to assign performance scores for each observed variable. Finally, scores are summed for each variable and converted into a percentage based on how many of the possible variables the officer performed. For example, in Table I an officer who “ends an encounter on a positive note” receives a six in Column A, “Weighted Score.” The total raw performance score for the encounter at the bottom of Column E is calculated as the sum of these scores. Then, the final performance score can be calculated by using the raw score as the numerator and the maximum possible potential score as the denominator:

\[ \frac{\sum N_{\text{scored behaviors}}}{\sum N_{\text{potential scores}}} \]

Thus, performance scores are expressed as a proportion of all behaviors that are possible in the encounter which are measured by the metrics.

It is important to stress that we do not expect that a perfect performance score is possible except, perhaps, in the least difficult encounters. NDM 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 of the important nuances of officer behaviors in encounters that can be used to gain a better understanding of police encounters with the public.

**Discussion**

This novel metric development process provides a practical method for rapidly creating tools for assessing individual officer performance while controlling for its relative difficulty. The process produced useful metrics for three types of common high-risk/high-consequence encounters that frequently require NDM: DFJDM, TSI, and CIT encounters. The time required to develop and employ metrics declined with experience from three years (DFJDM) to six months (TSI), to three months (CIT)[10]. We expect that the current metrics will require additional refinement, and that new metrics will be developed for other aspects of performance that require NDM.

When it comes to studying police encounters with the public, our method has substantial advantages over CTA – the other method for identifying measurable performance and difficulty variables in situations that require NDM. Not only is it less time intensive and expensive than CTA, it also produces interval-level metrics that can inform each stage of the often urgent problem-solving process facing police agencies. In time, evaluation, assessment, and refinement will ultimately determine the effectiveness of our metric.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weighted score</th>
<th>Applicable?</th>
<th>Potential score</th>
<th>Achieved?</th>
<th>Raw score</th>
<th>Performed (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td>Show signs of empathy</td>
<td>(-)</td>
<td>(-)</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>(-)</td>
</tr>
<tr>
<td>Apologize for inconvenience</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>De-escalate after being forceful</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>End encounter on a positive note</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Establish common ground</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>19</strong></td>
<td><strong>13</strong></td>
<td><strong>68</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** These data are used to calculate potential score, raw score, and percent performed (Columns C, E, and F, equations in italics)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weighted score</th>
<th>Applicable?</th>
<th>Potential score</th>
<th>Achieved?</th>
<th>Raw score</th>
<th>Performed (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td>Show signs of empathy</td>
<td>(-)</td>
<td>(-)</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>(-)</td>
</tr>
<tr>
<td>Apologize for inconvenience</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>De-escalate after being forceful</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>End encounter on a positive note</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Establish common ground</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>19</strong></td>
<td><strong>13</strong></td>
<td><strong>68</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table I.**

Example of TSI performance variables on a spreadsheet with weighted metrics score, applicability, and achieved score (Columns A, B, and D, entered)
development process for identifying and measuring what matters. But we think these types of metrics provide a solid starting place to understand problematic police encounters with the public that are likely to require NDM.

Following the “good-enough, quick-enough” maxim of police in urgent encounters, these interval-level metrics could be used now to address three of the most fundamental and elusive questions about police performance: “How much did new policies and training change officers’ behavior in the field?” “Did those changes have the desired impact on the problem?” And if not, “What should be done to get a better result?” This ability to empirically evaluate the effectiveness of policy, practice, and training makes it possible to refine any or all three until they appear to be satisfactory.

In a learning organization striving to meet the goals set by former President Barack Obama’s 21st Century Policing Task Force, this process likely would continue to evolve in order to adapt policies and practices in a changing world. Like biological evolution, these sorts of spiral adaptive approaches are very good at finding optima – “sweet spots” – where better outcomes are more likely in complex dynamic systems (Holland, 1995, pp. 41-90; Boehm et al., 2014).

The metrics we developed with this method provide a more plausible, extensive, and precise set of measures than were previously available – a set that can be validated, refined, and expanded over time. The ability to measure what officers do during encounters with the public gives us a clearer view of the critical juncture where state and citizen interact, services are provided, and justice flows or fails. We can no longer say, “That’s too complicated to measure.” At a minimum, we finally can begin to assess which of an officer’s behaviors during encounters with the public tend to increase the likelihood of a good outcome and which diminish it.

Limitations

Although this process for developing individual-level metrics appears to have great promise, several gaps in the research need to be addressed regarding the scales developed thus far. The following assessment of the metrics was conducted using Miller’s (1991, pp. 579-581) checklist of evaluative criteria for assessing a scale (pp. 579-581):

1. Item construction criteria: items developed during concept mapping arguably reflect the universe of important factors associated with the variables of interest and are simply worded. Item analysis by the true expert focus group indicates that they correlate with external criteria, the group’s go/no-go exercise eliminated items likely to be undesirable, ambiguous, relatively unimportant, or difficult to measure meaningfully.

2. Response set criteria: anonymity and random assortment[11] of scale items from scores of variables avoided common item rater bias problems (acquiescence, social desirability, desire to appear consistent, etc.). These bias sources also were controlled by the intensity of the focus group process and the true expert status of the participants.

3. Scale metric criteria: representativeness was addressed by careful selection of true experts from widely diverse and often divergent backgrounds, and by selecting expert item raters via online snowball samples seeded in multiple professional, training, and agency venues across the USA. Homogeneity was assessed by each focus group during concept mapping:

   a. Normative information about the meaning of responses cannot yet be obtained because there are no well-defined comparison groups at this time;

   b. Reliability was assessed among scale users in a laboratory setting, but not among item raters due to resource constraints. This confound may be countered
by the fact that the raters were experienced professional trainers and may be expected to have well-formed opinions; and

(c) Validity, the extent that the scales measure what they are supposed to measure compared to external criterion, has yet to be assessed.

We currently are analyzing results from two laboratory-based studies that will make it possible to begin validation. Ongoing use of the scales in field research and training also will feed into this assessment. As results are obtained from the use of the scales in various operational and research endeavors, it will be possible to begin assessing this issue.

In sum, studies to assess the gaps identified above in 3(a)-(c) are required. However, the majority of Miller’s evaluative criteria arguably have been met. We think it is reasonable to use the metrics in operational training and research settings given the absence of any other objective, empirically derived metrics for measuring police officer performance during encounters with the public. We note that both our TSI and CIT metrics have been used as the foundation of police and military training curricula. Also, we have demonstrated that they are efficient to employ once experienced scorers become familiar with the scoring items.

Implications and next steps

Science and innovation historian Steven Johnson (2014) emphasized that “when you have a leap forward in the accuracy of measuring something, new possibilities emerge”. As we discussed, one of the most fundamental things we do not know about police officers is how well they tend to perform day-to-day across a broad range of encounters with the public. One example of this critical gap has to do with police patrol officers – the government professionals in the USA with whom the public is most likely to interact. Patrol cops tend to have many encounters with people every day, often while working alone. The lack of impartial, objective information about the quality of officers’ performance leaves them, their supervisors, and the public in the dark. So we traditionally measure performance using surrogates such as general supervisory impressions, outcome measures such as productivity (e.g. numbers of arrests, calls handled, traffic tickets, etc.), and complaint/commendation data. This makes it difficult to identify gaps in an officer’s training and skill at managing encounters, or even assess the impact of that training on field performance. The question of whether use of the metrics to assess officer performance is better than existing methods is important.

Proposed validation experiments

Validation experiments are needed to determine whether the metric development process we describe produces metrics that are more complete, accurate, and precise than existing methods. A straightforward example of how such an experiment might be conducted follows below in brief. We would use a randomized, two-period crossover design to assess whether police trainers with experience on CITs who used the metrics to assess officers’ behavior in encounters predicted the probability of a desired outcome significantly better than those who made such predictions using only expert observation. The second period of this design also would test whether learning to employ the metrics and practicing with them might improve participants’ expert insights when practice effects and exposure to the treatment stimuli were controlled for. This would address the issue of their utility for training design and gap analysis. Figure 2 illustrates the design flow.

Period 1: recruit 200 police trainers with experience on CITs as participants, then randomly assign them in Period 1 to equal-sized treatment (Trt A) and control (Trt B) groups. Trt A participants would receive the CIT metrics and be trained up to criterion level in their use to score officer performance. Each of them would separately view video footage of 100 different, randomized CIT encounters that previously had been evaluated
for difficulty, and for which outcomes had been blinded. Participants would then review each video and score officer performance using the metrics (scenario score range = 0-100). Trt B participants would not have access to the metrics, but they would be asked to separately review the same footage, and subjectively rate overall officer performance in each scenario by assigning a score from 0-100. Analyses of Period 1 results would test whether: there were significant differences between Trt A and Trt B scores for performance; and which method better predicted the outcome. This would assess: whether the metrics improved officers’ ability to predict encounter outcomes; and whether the metric’s detailed scales appeared to be more effective for measuring what matters in encounters.

Period 2: Trt A participants cross over to the control condition and Trt B participants cross over to the treatment condition. Otherwise using the same protocol as in Period 1, Trt A participants would review, and then score a re-randomized set of the scenarios without using the metrics, and Trt B participants would receive the metrics training and then use them to evaluate performance. If Trt B participants were better able to predict outcomes with the metrics in Period 2, it would support the hypothesis that the metrics improve officers’ ability to predict encounter outcomes – even after previous exposure to the same of stimuli in a different order during Period 1. If Trt A predictions without the metrics worsened, that also would support the hypothesis that the metrics improve officers’ ability to predict encounter outcomes. Conversely, if Trt A predictions without the metrics improved in Period 2, it might suggest that using the metrics in Period 1 had improved their expertise – but only if increased exposure to the stimuli and practice effects had been controlled for by Trt B participants having benefited from the metrics in Period 2 despite prior exposure to the stimuli.

Future research
We think that the most elusive and important question that this method could be used to evaluate is whether changes in policy, practice, and training promote changes in officer behavior during encounters with the public. The following examples suggest research that could be conducted using the performance and difficulty metrics we have developed for DFJDM, TSI, and CIT:

- experimental evaluations of the causal link between deadly force rules of engagement, accountability systems, or training and real world performance;
- experimental evaluations of the impact of different work-hour practices and officer fatigue on performance in real and simulated use-of-force encounters;

Figure 2.
Two-period, two treatment (2 × 2) crossover study design

Notes: In Period 1 treatment group A (Trt A) participants receive the treatment by using the metrics to assess officer performance and treatment group B (Trt B) participants act as controls by making assessments subjectively. In Period 2, Trt B participants receive the treatment and Trt A participants act as controls.
exploring individual characteristics (e.g. experience, training, risk-tolerance, cognitive abilities, etc.) that may predict how an officer will perform in a deadly encounter;

exploring the utility of the metrics for measuring team performance among members of various sizes and types of operational units; and

experiments to determine whether a perfect performance score is possible, and assess the distributions of performance scores for different situations based on both laboratory and field observations, while controlling for difficulty.

Conclusions

This method for developing metrics to measure what matters in encounters between police and the public can have both empirical and political advantages. Politically, the metrics can be used to help focus attention on causal processes, not just tragic outcomes when rhetoric and outrage feed impatience and devalue evidence and logic. Our method can be used as an impartial, science-based technique for quickly mediating conflict by focusing attention on a research question:

\textit{RQ1. What do we know, what do we need to know, and what can be done about it here and now?}

This acknowledges that policies, training, and interventions are hypotheses that must be tested and refined – but also that real world crises cannot always wait for gold-standard randomized control trials.

New possibilities are created by metrics that measure what officers do at the micro level and assess the relative difficulty of each encounter. Chief among them, we think, is that they can help build trust and nurture legitimacy on both sides of the police-citizen divide. Balancing justice in the streets with justice in the workplace for police requires accurate and objective measurement. Officers must be held accountable for their actions – but our expectations for their behavior must reflect the limits of human performance, the unpredictability of encounters, and the validity of the training they receive.

Acknowledgments

The authors have no conflicts of interest to report. Funding from the following organizations supported this research. However, the information reported here and the conclusions and opinions offered are those of the authors alone and do not necessarily reflect the funding agencies and organizations’ policies, practices, or opinions:

- “Critical Job Tasks Simulation Laboratory Expansion for WSU Sleep & Performance Research Center.” Office of Naval Research (Grant No. N000140810802).
- “Impact of Work-Shift Related Fatigue on DFJDM, Driving, Cognition, and TSI Performance.” Office of Naval Research (Grant No. N0001411110185).
- “Empowering the Strategic Corporal: Training Young Warfighters to be Socially Adept with Strangers in Any Culture.” Defense Advanced Research Projects Agency (Grant No. W911NF-12-0039).
- “Developing Objective Interval-Level Metrics to Measure the Effectiveness of CIT Training.” Spokane (Wash) Police Department (Contract No. OPR 2013-20130455).

The authors acknowledge the active participation of Lauren B. Waggoner during the DFJDM metric development process, and Elizabeth J. Dotson with scoring scenarios and writing assistance. Cynthia L. Morris provided invaluable critical advice. The authors also acknowledge the scores of colleagues, students and outside experts who advised and assisted in the DFJDM, TSI, and CIT projects and regret that there is not enough room to name them here. The authors are deeply grateful to the true experts who participated in our scale-development processes and to the more than 1,000 police trainers from across the USA who scored the scales.

Notes
1. The primary focus of this paper is on the metric development process, its potential utility, and policy implications. However, illustrative examples from the metrics themselves are used as extensively as space constraints allow.
2. Roughly nine different NDM models were conceived during the first years of study in this area (Klein, 2008). All are sufficiently similar for the purpose of this paper (Lipshitz, 1993).
3. Building trust and legitimacy, policy and oversight, technology and social media, community policing and crime reduction, training and education, and officer wellness and safety.
4. The three sets of metrics are published in final report form (see Acknowledgments and References), and are available upon request.
5. This is a common phenomenon among true experts because their expertise grows organically as experience and insight accrete (Shanteau, 2001).
6. These projects were approved by the authors’ Institutional Review Board.
7. The TSI metrics were developed for the Department of Defense, and therefore needed to be relevant to both military and police personnel.
8. For mathematical and experimental proofs of why this process yields interval-level data from ordinal-level expert rankings, see Thurstone (1959, Chapters 18 and 19).
9. This practice is consistent with Thurstone’s (1959) discussion of units of measure (pp. 223-225).
10. Note that CIT was developed for a community after the in-custody death of a developmentally challenged person which was caused by officers’ mishandling him when they responded to an erroneous robbery-just-occurred call. Speedy development of CIT metrics helped defuse community outrage, quickly establish a multi-agency training program, and educate practitioners and the public.
11. Random assortment of scale items was discontinued after DFJDM. The overhead associated with randomization of roughly 800 items and managing the online survey instruments was high, and rater debriefings indicated that they found the constant reappearance of such similar items interfered with providing coherent responses. Several raters said it made it more difficult to characterize scales with non-linearities.

References


Further reading


Corresponding author

Bryan Vila can be contacted at: vila@wsu.edu

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com