Policy Notes

'Remodelling:' The Need for More Robust Models and Metrics for Counterterrorism Threat Analysis

by Jason A. Bakas

Abstract

In this Policy Note the author addresses shortcomings found in many current proprietary counterterrorism threat assessment tools used by government agencies. In addition, he provides evidence which makes a strong case for the adoption of a Structured Professional Judgment methodology, to be used as the basis of future proprietary or in-house terrorist threat assessments within law enforcement and intelligence agencies.

Keywords: actuarial, counterterrorism, intelligence agency, law enforcement, structured professional judgment, terrorist threat assessment, validity

Introduction

Terrorism threat assessments are important tools used throughout law enforcement and intelligence agencies internationally. These assessments are designed to evaluate the threat specific terrorist groups and lone actors pose. The models are necessary for counterterrorism (CT) agencies to be able to minimize poor judgments in the form of both false positives - that is identifying individuals or groups as a terrorist threat, when in fact they are not intending to engage in terrorist acts - and false negatives - where individuals or groups who are intending to commit acts of terrorism are not considered dangerous, which may result in a terrorist attack being committed. In many cases, organizations responsible for CT develop their own proprietary threat assessments. This is often done to ensure the threat assessment fulfils the operational needs related to a counterterrorism organization's abilities. However, there are some critical issues that have been identified with regards to many of these *in-house* terrorist threat assessments. These issues are not only theoretical, but have been witnessed by the author of this Policy Note as critical flaws in a number of proprietary CT threat assessments internationally.

Questionable Validity

One of the main concerns found with many proprietary threat assessments is questionable validity. A critical process in evaluating the effectiveness of any threat assessment metric or model, is to evaluate it for both reliability and validity. These frameworks are used to determine if an assessment instrument is actually measuring what it claims to be measuring, and that the resulting judgements that come from employing the assessment instrument accurately reflect the outcomes of what is meant to be assessed. To paraphrase renowned researchers and assessment developers Douglas and Kropp - without evidence of reliability and validity, a threat assessment instrument is valueless.[1] This author has found that rigorous tests of validity and reliability are rarely conducted or reported. This widespread tendency to under report - or not conduct - tests of validity and reliability leaves many proprietary CT threat assessments in a precarious situation. It begs the question: *Is it simply under-reporting or hidden invalidity? We just don't know.* When encountering these proprietary threat instruments, the author would enquire about how the assessment was developed, its validity and how the validity was tested. These questions would be directed to the sub-units or research divisions which created the assessments. In almost all cases, the question of appropriate constructs and validation testing was not answered, or poorly answered. This leaves it unclear how these metrics were developed; how indicators were defined; the degree of efficacy the assessment has; or if the metric is based on robust empirical or theoretical evidence.

Further, in many cases where CT agencies did conduct testing on their proprietary threat assessment models, researchers placed an over-reliance on *Cronbach's α,* as the sole source of evidence demonstrating validity and reliability. The problem here is that these assessments may seem to demonstrate strong structural validity on the surface, but when subjected to more rigorous tests beyond Cronbach's α, they may have significant shortcomings. These more rigorous tests include examining for scores across time with *Test-Retest Reliability*; or the structure of the latent constructs with *Confirmatory Factor Analysis;* or *Measurement Invariance* for an assessment's equivalence across populations; or calculating *Cohen's Kappa Coefficient*, *Kendall's Coefficient of Concordance*, or *Intraclass Coefficient* for inter-rater reliability. Very few of the CT assessments seen by this author were examined with any of these more advanced methods. This raises some serious questions – the first being: *why not*?

You Don't Know, What You Don't Know

The answer to a lack of testing might come from a 2017 study conducted by Flake and colleagues, which found that tests of structural validity, such as measurement invariance, are poorly understood and infrequently conducted in many scholarly works within social and personality measurement psychology (the structure of psychometrics are very similar to threat assessments, they are both latent variable models).[2] Moreover, Flake and colleagues study found many of these same scholarly works rarely reported rigorous methodologies for testing validity or these lacked appropriate validation testing completely. In addition, a 2008 study by Aiken and colleagues, found that measurement and test theory is often ignored in doctoral level graduate studies within psychology and that only a minority of doctoral students know how to apply the methods of reliability testing correctly.[3] Thus, if many doctorate level academic researchers lack this knowledge, we can make the inference that many practitioner researchers may also lack the knowledge in measurement and Classical Test Theory. However, this is a poor excuse. Practitioner researchers and their instruments should be held to the highest standard. *Lives are depending on them getting it right*.

The Issues with Actuarial Metrics as Terrorist Threat Tools

The lack of proper reliability, and validity (weather it be based on a lack of knowledge or otherwise) brings to light a much larger issue regarding the efficacy of many proprietary CT threat assessments. Because many organizations don't test - or don't know how to properly test their assessments - we can not see whether there are some fundamental problematic issues that are underlying the construction of these metrics. The vast majority of these proprietary threat assessments witnessed by the author (almost all), are what can be defined as **actuarial models** – they are latent variable models that employ the use of a '*check list*' system of fixed numerically weighted indicators.[4] These indicators are scored, using statistical formulas to calculate and conclude a predictive threat score on a numerical scale, or as a predictive percentage. The advantages of actuarial metrics include their ability to allow objectivity in decision-making and a high inter-rater reliability across evaluators. However, in this author's opinion, there are some critical limitations which make actuarial metrics highly problematic for terrorist threat assessment. This is especially the case when they are constructed by researchers who themselves have a poor understanding of measurement - and test-theory. The main issue here is the tool's generalizability across terrorist actors. Actuarial applications for assessing terrorist threats lack strong invariance. They may work well for a specific type of terrorist actor or group who remain static, as unchangeable entities - but they do not work so well when applied to a spectrum of terrorist individuals, or groups of individuals; or the same terrorist individuals or groups of individuals in different contexts or settings; or over time. Such a statement is admittedly controversial and may raise more than a few eye brows – but as explained below in further detail, this should become more evident.

Poor Linguistic Invariance

There are hundreds of definitions of terrorism, and equally hundreds of scholarly works which discuss the various definitional issues of terrorism. For example, in 2011 Easson and Schmid identified 260 definitions of terrorism that hold a level of validity in describing this phenomenon.[5] Some researchers and analysts hold that terrorism must be motivated by ideology, while others state it can (also) be motivated by personal or vicarious revenge, or other idiosyncratic incentives.[6] Many other examples of disagreement exist, e.g. regarding raicalization or extremism. The same definitional disagreement issues are found within government and CT organizations, which makes sharing intelligence data sometimes potentially problematic. This issue becomes even more complicated when it comes to additional definitional issues, such as those surrounding 'lone wolf' or 'lone actor' terrorism. For example, some agencies include in their monitoring individuals with severe mental illness who are inspired or directed to engage in attacks on behalf of an extremist movement. Others exclude these, because mental illness-related lethal violence is considered to be 'mass murder' and not terrorism, even if it is linked to a larger extremist movement. The boundaries of terrorism are inevitably fuzzy, and if we do not have a precise way of defining terrorism - then we really do not know what exactly is - or should be - considered a threat indicator for terrorism. Because of this, we do not know the extent to which an actuarial model includes all the right elements, and excludes all the irrelevant elements. This issue becomes more problematic when assigning weights or numerical values to these indicators.

Questionable Indicators

In relation to questionable concepts, and the notion of *what exactly is - or should be - considered a threat indicator for terrorism,* the author has found that many threat assessment instruments incorporate empirically questionable indicators – indicators that may have been found to be related to threats in some isolated cases, but are not necessarily applicable across most individuals or groups of terrorists. Within academia, the issue of a lack of robust generalizable indicators to terrorism propensity is rooted in low base-rates.[7] Base-rates are statistics used to describe the percentage of a population that demonstrates some characteristic. The issue of low base-rates in scholarly works can be the result of a relatively small body of empirical data, which stems from the obvious challenges of academics having access to – and being able to publish – operational terrorist data acquired from CT organizations. Despite CT organizations housing a large body of empirical data (which is acquired in the course of their investigations) – *we do not really know if concepts found in the 'in-house data' used to develop indicators are truly generalizable*. This is due to a couple of reasons. The first being, the type, scope and quality of data collected. If we look at the way the UK's Metropolitan Police has collected data on suspected gang members, we can see how datasets held by law enforcement organizations could be problematic.[8] A May 2020 report by Amnesty International found that the Metropolitan Police's Gangs Matrix - a database of suspected gang members in London designed to be used by police to prevent serious gang violence – had collected data in a "chaotic, inconsistent" manner and was "not fit for [its] purpose".[9] Reports state that the threshold for data collection was "very low" with "no clear guidance or criteria, and wide discretion for police officers and partner agencies".[10] If this same problematic issue exists also within CT organizations, there would likely be challenges in effectively operationalizing data. The second issue, as previously stated, is a lack of appropriate construct and validation testing. *Because we have seen an indicator's relevance in a limited number of 'N' cases – does this mean it is relevant across a spectrum of most terrorist actors? We don't know* – and without proper testing we cannot make claims about the generalizability of indicators with a high level of confidence. Moreover, because terrorist investigations and terrorist groups or individuals are nuanced in situational and dynamic contexts, it raises the question – *assuming we had well collected data, what are the chances of finding robust generalizable indicators, that would be efficacious, when applied in actuarial models?*

By definition, an indicator must vary systematically with changes in the latent construct - it must increase or decrease monotonically with that latent construct. In other words, when higher or lower scores are observed on the indicator, this must be related to an increase or decrease in the latent construct's values. However, many of the indicators found by this author in proprietary threat assessments have been shown to not be generalizable within

the academic literature – and, as stated, there is no evidence to suggest they have demonstrated generalizability in 'in-house' CT datasets. For example, the author has seen "*time spent consuming violent extremist media*" as a threat indicator on more than one nationally used terrorist threat assessment. However, those who are found to be both consumers and producers of violent extremist media are not necessarily on a trajectory for engaging in terrorism violence.[11] The same can be said for radicalization. "*The degree or severity of radicalization*" was found to be present in many proprietary actuarial threat assessments, but it has been well documented that holding an extremist or radical ideology does not necessarily put an individual on a trajectory for engaging in terrorism violence. In fact, most persons who hold an extremist ideology do not themselves engage in violence.[12] Moreover, not everyone who engages in violence, in the name of a terrorist group or movement, holds an extremist or radical ideology.[13] Therefore, the best evidence available demonstrates that indicators such as extremist media consumption and radicalization, do not necessarily vary monotonically with the construct of interest - which in our case is the threat of engaging in a terrorist attack. Thus, models that use this type of indicators are making an indirect inference. If we do not have accurate threat indicators, how can we expect to predict or measure the threat of terrorism? This begs the question: *why are we selecting imperfect indicators for assessments and mechanically apply them the same way every time?*

While all types of assessment instruments are reliant on the assumption of validity generalization from population data to an individual - this is especially the case in actuarial models, because these have a great assumption burden, due to the fact that they function mathematically or algorithmically.[14] Thus, for actuarial models to be efficacious we need probability statistics (i.e. base-rates) of a sufficient breadth of quality and quantity. *But is this realistic*? As stated, each terrorist investigation and terrorist group or individual is nuanced in situational and dynamic contexts. The weight of indicators may need to be based upon case-specific details. Actuarial fixed weighted indicators calculated with an algorithm are not going to be able to account for this. As a consequence, in some cases they may pay too much attention to certain indicators and not enough (or ignore) to other indicators. Because it is mathematical or algorithmic, the process of how threat predictions were made cannot be reviewed – we just have to trust the algorithm. It is highly problematic to use imperfect fixed weighted indicators to make threat predictions in cases where we know a dynamic mixture of nuance, situation and context matter. On top of this, we cannot check the process to make sure it is accounting for and evaluating the most important factors. We are just blindly trusting the assessment even though we know indicators may be flawed - *that's like trusting your wrist watch, when it doesn't tell time all that well and when time might be changing.*

Poor Measurement Invariance

The issue of poor generalizability of indicators, brings to light the issue of assessment measurement invariance. Measurement invariance is a scaled capacity to measure the same construct in a comparable way across populations or across contexts. In other words, it is the degree of generalizability of an evaluation. Outside of terrorism, this is often tested to see if an assessment or metric is well represented across different genders or cultures. The question of measurement invariance (assuming concepts are defined in a consistent manner) can be answered by applying statistical evaluations, such as the application of *Multi-Group Confirmatory Factor Analysis.* However, the vast majority of actuarial proprietary CT threat assessments encountered by this author were unable to demonstrate strong measurement invariance - yet each assessment claimed it. Again, this widespread tendency to not demonstrate - or not conduct – appropriate testing leaves many proprietary threat assessments in a precarious situation, leaving us with the question: *do they really work the way they say they work?*

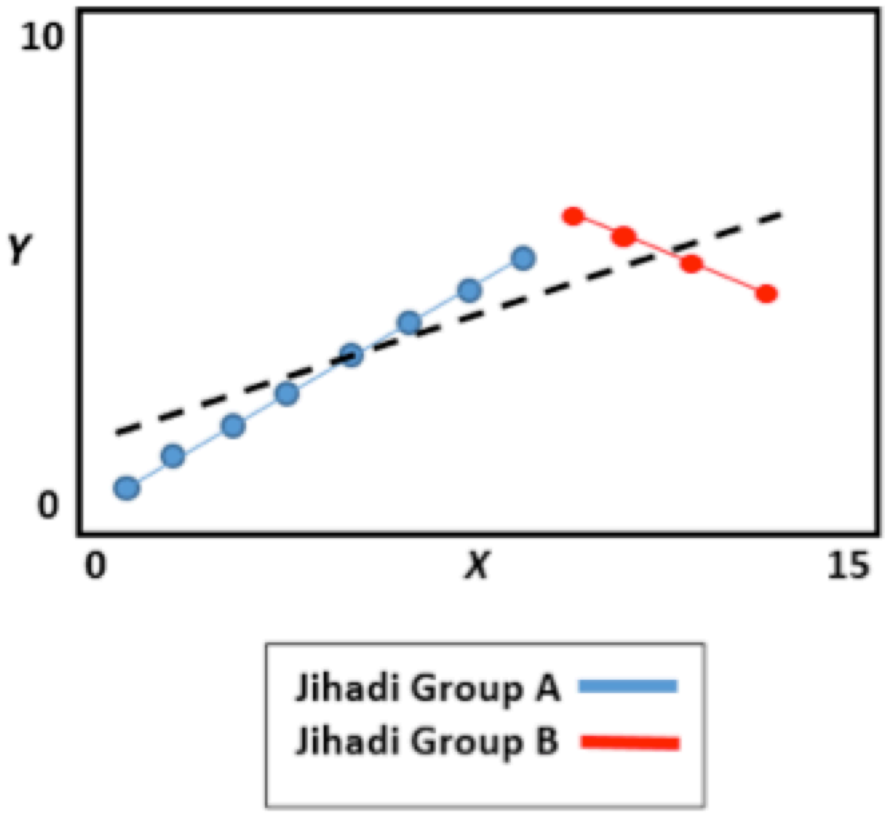
Issues related to linguistic invariance and questionable indicators, create inherent problematic issues with measurement invariance. This is especially the case in actuarial models, as stated previously, since these models hold a great assumption burden, due to the fact that they function statistically.[15] If an assessment developer uses a small population data set of a specific terrorist type or sub-type as his or her representative sample, we are likely to see qualities of validity and reliability only hold in relation to the given test population. The

generalizability to other terrorist populations, or even the same terrorist population were changes occur, is questionable. *What's true for jihadists might not be true for white nationalists, ethnic secessionists, or even for a different jihadist group.* If this lack of generalizability was related to genders or cultures, these assessments would be seen as biased towards a gender or cultural majority. In its application to terrorism, this is still a bias, but bias towards the specific terrorist type or sub-type sample used in the development of a measurement instrument. In many cases assessment developers in CT organizations are not testing the relation between test scores and criterion variables or outcomes, to see if they are consistent across groups, using large samples. They are simply just assuming strong invariance. These assessments are then rolled out with false confidence, which may lead to an intelligence or security agency adopting and implementing a potentially flawed assessment.

Simpson's Paradox

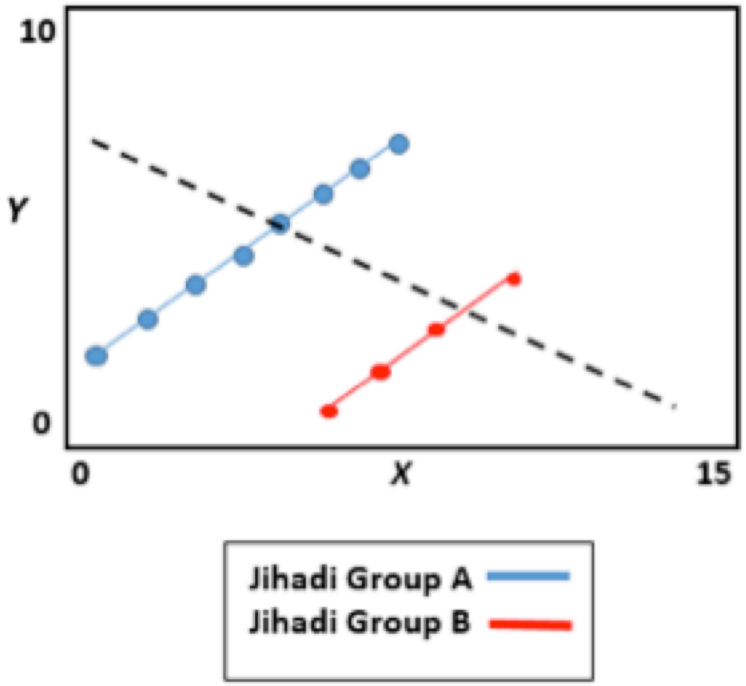
Another, and most problematic issue with the use of actuarial metrics for proprietary threat assessment is Simpson's paradox. This is a phenomenon in which individual trends appear in different groups of data but disappears or reverses when these groups are combined. For example, assuming the indicators are accurate, if we apply a high trending threat group of individuals – '*Jihadi Group A*' (who are consistent with the data set majority used in development) - to a threat assessment, we would likely see a relationship were the greater the presence or degree of indicators, the greater the threat outcome. If we also plotted a high – yet low trending threat group of individuals – '*Jihadi Group B*', on its own, we see a trend in the opposite direction. However, if we do not separate the groups, we would see the regression line in the direction of '*Jihadi Group A*'. Therefore, a CT agency would consider individuals in '*Jihadi Group B*' as moving towards a high threat direction, when in fact it is not (see Figure 1).

Figure 1:



We might even see a positive trend (high threat) in both groups when they are separate, but when the groups are combined, we see a negative trend; resulting in a low threat (see Figure 2).[16]

Figure 2:



We can see from this, that in many cases actuarial threat metrics do not always demonstrate an accurate threat picture. The issue is, we need to know when to break down our terrorist populations into groups. However, this author has encountered many proprietary threat metrics that are meant to be used as universal for all terrorists - or for all terrorists of a specific subtype. *So why would a CT analyst identify distinct factors and collect for group differences if they are told they do not need to?* It is important to note, that no test model is capable of perfectly capturing the theoretical variables of interest and that some degree of *error* is unavoidable, but this can be accounted for with proper statistical test analysis. The issue here is with *Systematic Error* sometimes called *Statistical Bias* – it is bias built into the test. This is not a question of inaccurate intelligence leading to a faulty threat picture, as is often claimed. Rather it is the reliance on threat assessment instruments that have not been properly developed or evaluated before they are rolled out for application.

One Solution may be Adopting an SPJ Approach

The Structured Professional Judgment or SPJ method of assessment has been used in medical practice and psychology for decades. In fact, in recent years a handful of these SPJ psychometric violence assessments have been designed by academics for use in evaluating terrorism violence – primarily to be utilized in correctional settings.[17] Despite this trend, in all the in-house developed CT threat assessments this author has seen, few used the SPJ approach.

Like actuarial metrics, the SPJ model provides assessors with a structured *'check list'* of indicators that are rooted in strong empirical or theoretical evidence. However, it differs as this '*check list*' does not hold the same level of rigidity. Rather, the list of indicators functions as a set of systematic guidelines for evaluating the outcome construct - in our case terrorism threat. The SPJ methodology '*unlocks*' the indicators from fixed numerically weighted rankings and allows the evaluator (in our case a CT analyst or case officer) to attribute the presence and relevance of indicators - through interpreting and appreciating information related to the threat propensity of the individual or group under investigation. Therefore, unlike actuarial metrics, which may pay too much attention to certain indicators and not enough or altogether ignore other indicators, the SPJ method allows the evaluator the flexibility to take into account situational and contextual factors, while still

being guided in decision-making with a structured approach.

As a result of this flexibility, the SPJ method mitigates against issues discussed as problematical in this Policy Note, such as poor invariance and test bias. Further, these SPJ models are not predictive but rather probabilistic, and thus attribute the presence or degree of indicators to an increased level of threat, not a determinate level of threat. As mentioned before, SPJ tools have been used in physicians' medical assessments for years. For example, while conducting a health check-up, if a doctor were to find a patient to have (i) hypertension (ii) be a smoker; (iii) suffer from obesity; (iv) have a family history of heart disease and (v) have diabetes - that physician would be able to conclude that the patient has an increased probability for cardiovascular failure -related death (i.e. all the major indicators are present). However, if only one or two of those indicators are present (lets say hypertension and family history) an individual could be potentially at the same level, or higher level, of probabilistic risk of cardiovascular related death. As we know, not everyone with all heart disease indicators will have a heart attack while some people with only one or two of those indicators will actually have a heart attack and die as a consequence. This same principle applies to terrorist threat. It is not necessarily the number of indicators, but rather the value of those indicators found to be present within the context of the case that matters. Because of the SPJ method's flexibility, and its design to allow the evaluator to interpret and appreciate information, it has been argued that the SPJ approach is the more appropriate and the most fruitful form of assessment when completing evaluations under conditions where the information available is limited and often is also of poor quality.[18] This is often the case during active investigations, when time-pressured CT professionals do not have enough data on-hand that would meet a 'clinical standard.'

While most familiar with SPJ maintain that this flexibility is a strength of the method, others have criticized it as a weakness. It has been suggested that the interpretive nature of the methodology can lead to decision-making bias, and even that the SPJ approach is ill-suited for law enforcement or other security agency use. However, research conducted by Storey and colleagues found that, following proper training, SPJ assessments could be accurately used by police and other criminal justice professionals.[19] Moreover, many scholars argue through training and acquiring a background knowledge on indicators to violence and the population type being assessed, issues of bias can be minimized or even off-set. *When it comes to CT, anyone applying or using any type of threat model should be highly knowledgeable about terrorism in general and the terrorist group or individual they are assessing.*

There Is Still the Problem of Questionable Validity

Of the few SPJ proprietary CT threat assessments this author had the privilege to observe, many did not provide strong evidence – or any evidence - related to appropriate construct and validation testing. *This takes us back to the same issue* - it is unclear how these metrics were developed; how indicators were defined, the degree of efficacy the assessment has; or if the metrics used are based on robust empirical or theory-driven evidence. Even though SPJ assessments do not apply algorithmic or statistical computations in determining threat attribution – they still must be developed and tested with strong scientific rigor. These models still need to demonstrate reliability and validity; indicators still must demonstrate evidence of a relationship-based outcome with the latent construct; and the indicators must still demonstrate evidence of generalizability to the terrorist population of interest. Oversights in appropriate testing could lead to false positive or false negative assessments of threats, which could be significantly detrimental to public safety. In short, reliability and validity testing must be conducted, and should be reported.

Conclusion

In this Policy Note, a number of shortcomings found in many proprietary threat assessments that have been developed by law enforcement CT organizations were presented. In the author's opinion, many of these arose from a poor understanding of model development and evaluation testing. These issues are amplified when

untrained or unskilled model developers attempt to create functional actuarial threat metrics. Because terrorism is an elastic and amorphous concept, the author argues that - even at best - CT actuarial threat assessments are problematic, and he concludes that the adoption of a Structured Professional Judgment methodology would likely be more efficacious. Therefore, it is recommended, that the Structured Professional Judgment method should be adopted as basis of future proprietary or in-house terrorist threat assessments. Now more than ever, forecasting the threat terrorists, or potential terrorist actors, pose is of utmost importance; we cannot afford to get it wrong given the high number of potential perpetrators. A 2020 report by the UK government found that British intelligence agencies are aware of more than 43,000 individuals who pose a potential terrorist threat to the UK.[20] Of that number, 3,000 are considered '*subjects of interest*' and are under active investigation. If CT agencies are going to develop their own threat assessment metrics and models, they need to have all the knowledge to get it right. To paraphrase Victoroff – a lack of good understanding on terrorism has left many CT policy makers to design counterterrorism strategies without the full benefit of facts – or – worse - be guided by theoretical presumptions assumed to be factual.[21]

Disclaimer: The views expressed in this Policy Note are the author's and the author's alone. They do not necessarily reflect the opinions of the author's professional or academic affiliations. The counterterrorist threat assessments discussed in this Policy Note have been personally examined by the author. The names of the agencies which develop and own the assessment instruments discussed here, the number of assessments examined, as well as the nature and contents of these assessment tools, can, unfortunately, not be disclosed here for security reasons.

About the Author: **Jason A. Bakas** *is an intelligence professional and researcher. He has been recognized internationally for his work in advancing risk and threat assessments in the application of counterterrorism and counter-organized crime. He holds a Master's of Arts from American Military University's School of Security and Global Studies and was a student of the National Consortium for the Study of Terrorism and Responses to Terrorism (START), at the University of Maryland. Correspondence ought to be addressed to: Jason.Bakas@consultant.com*

Notes

[1] Douglas, K. S., & Kropp, P. R. (2002). 'A prevention-based paradigm for violence risk assessment: Clinical and research applications'. *Criminal Justice and Behavior, 29*(5), 617–658; URL: https://doi.org/10.1177/0093854023236735

[2] Flake, J. K., Pek, J., & Hehman, E. (2017). 'Construct validation in social and personality research: Current practice and recommendations'. *Social Psychological and Personality Science*, 8(4), 370-378; URL: https://doi.org/10.1177/1948550617693063

[3] Aiken, L. S., West, S. G., & Millsap, R. E. (2008). 'Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America'. *American Psychologist,* 63(1), 32–50; URL: https://doi.org/10.1037/0003-066X.63.1.32

[4] See the Federal Bureau of Investigation's (FBI) *Indicators of Mobilization to Violence* (IMV), as an example of a counterterrorism actuarial model which employs a check list system of fixed numerically weighted indicators. *Please note:, neither this Policy Note, nor the author are stating, implying or otherwise suggesting any factors discussed in this paper are in any way related or relevant to the IMV*. The IMV is simply cited as an example of a counterterrorism actuarial model used in law enforcement. The IMV was leaked to the public in 2017: https://www.documentcloud.org/documents/3460923-Imv-Score-Final.html#document/p1

[5] Easson, J.J., & Schmid, A. P. (2011), '250+ Academic, Gvoernmental and Intergovernmental Definitions of Terrorism'; in: Schmid, A. P. (Ed.). *The Routledge Handbook of Terrorism Research*. New York and London: Routledge, pp. 99-200.

[6] Khalil, J. (2014). 'Radical beliefs and violent actions are not synonymous: How to place the key disjuncture between attitudes and behaviors at the heart of our research into political violence'. *Studies in Conflict & Terrorism*, *37*(2), 198-211; URL: https://doi.org/10.1080/1057610X.2014.862902; Spaaij, R. (2011). *Understanding Lone Wolf Terrorism: Global Patterns, Motivations and Prevention*. Dordrecht: Springer Science & Business Media. p. 856.

[7] Gill, P., Horgan, J., Corner, E., & Silver, J. (2016). 'Indicators of lone actor violent events: The problems of low base rates and long observational periods'. *Journal of Threat Assessment and Management, 3*(3-4), 165–173; URL: https://doi.org/10.1037/tam0000066; Herrington, V., & Roberts, K. (2012). Risk assessment in counterterrorism. In U. Kumar & M. K. Mandal (Eds.), *Countering Terrorism: Psychosocial Strategies* (pp. 282–305). London, United Kingdom: Sage; Pressman, D. E., & Flockton, J. (2014). Violent extremist risk assessment: Issues and applications of the VERA-2 in a high-security correctional setting. In A. Silke (Ed.), *Prisons, Terrorism and Extremism: Critical issues in management, radicalisation and reform* (pp. 122–143). London, United Kingdom: Routledge; Sarma, K. M. (2017). 'Risk assessment and the prevention of radicalization from nonviolence into terrorism'. *American Psychologist*, *72*(3), 278–288; URL: https://doi.org/10.1037/amp0000121

[8] Of note, neither this Policy Note, nor the author are stating, implying or otherwise suggesting that any counterterrorism organization collects data in an improper or inaccurate way, or in a manner that is discriminatory or racialized. The comparison to the way the UK Metropolitan Police collected data is simply for reference purposes, to provide the reader with an open source example of how police organizations may have erred in data collection.

[9] Amnesty International UK (2020, May 18). *Trapped in the Gangs Matrix*. URL:https://www.amnesty.org.uk/london-trident-gangs-matrix-metropolitan-police; Full Report: Amnesty International. (2018). *Trapped in the Matrix: Secrecy, stigma, and bias in the Met's gangs database*; URL: https://www.amnesty.org.uk/files/2018-05/Trapped%20in%20the%20Matrix%20Amnesty%20report.pdf?lJSxllcKfkZgr4gHZsz0vW8JZ0W3V_PD=

[10] Ibid.

[11] Brachman, J. M. (2010 Jul, 29). 'My Pen Pal, the Jihadist'. *Foreign Policy*. URL: http://foreignpolicy.com/2010/07/29/my-pen-pal-the-jihadist/; Brachman, J. M. (2010 Oct, 10). "Watching the Watchers." *Foreign Policy*.URL: https://foreignpolicy.com/2010/10/12/watching-the-watchers/21; Brachman, J. M. (2008). *Global Jihadism: Theory and Practice*. New York:Routledge.

[12] Borum, R. (2011). 'Radicalization into violent extremism I: A review of social science theories'. *Journal of Strategic Security*, *4*(4), 7-36; URL: http://dx.doi.org/10.5038/1944-0472.4.4.1; Horgan, J., & Taylor, M. (2011). 'Disengagement, de-radicalization and the arc of terrorism: Future directions for research'. In R. Coolsaet (Ed.), Jihadi Terrorism and the Radicalization Challenge (pp. 173-186). London, UK: Ashgate.

[13] Borum, R. and Robert Fein (2017) 'The Psychology of Foreign Fighters', *Studies in Conflict and Terrorism* (40) 3, 248-266; URL: https://doi.org/10.1080/1057610X.2016.1188535; Holbrook, D., & Horgan, J. (2019). 'Terrorism and Ideology: Cracking the Nut'. *Perspectives on Terrorism*, *13*(6), 2-15; URL: https://www.universiteitleiden.nl/binaries/content/assets/customsites/perspectives-on-terrorism/2019/issue-6/01-holbrook-and-horgan.pdf; Khalil, J. (2014). 'Radical beliefs and violent actions are not synonymous: How to place the key disjuncture between attitudes and behaviors at the heart of our research into political violence'. *Studies in Conflict & Terrorism*, *37*(2), 198-211; URL: https://doi.org/10.1080/1057610X.2014.862902; Neumann, P. R. (2015). 'Victims, Perpetrators, Assets: The narratives of Islamic State defectors'. King's College London: ICSR, p.9; URL: https://icsr.info/wp-content/uploads/2015/10/ICSR-Report-Victims-Perpetrators-Assets-The-Narratives-of-Islamic-State-Defectors.pdf .

[14] Hoekstra R, Kiers HAL and Johnson A. (2012) 'Are assumptions of well-known statistical techniques checked, and why (not)?' *Front. Psychology* **3**:137; URL: https://doi.org/10.3389/fpsyg.2012.00137

[15] Ibid.

[16] Clifford H. Wagner (1982) 'Simpson's Paradox in Real Life', *The American Statistician*, 36:1, 46-48; DOI: 10.1080/00031305.1982.10482778; an example of Simpson's paradox: a positive trend appears for two separate groups and a negative trend appears when the groups are combined. URL: https://en.wikipedia.org/wiki/Simpson%27s_paradox#/media/File:Simpson's_paradox_continuous.svg

[17] Lloyd, M. (2019). 'Extremism Risk Assessment: A Directory'. *The Centre for Research and Evidence on Security Threats (CREST)*; URL: https://crestresearch.ac.uk/resources/extremism-risk-assessment-directory/

[18] Sarma, K. M. (2017). 'Risk assessment and the prevention of radicalization from nonviolence into terrorism'. *American Psychologist*, *72*(3), p.280; URL: https://doi.org/10.1037/amp0000121

[19] Storey, E. J., Gibas, A. L., Reeves, K. A., & Hart, S. D. (2011). 'Evaluation of a violence risk (threat) assessment training program for police and other criminal justice professionals'. *Criminal Justice and Behavior, 38,*554–564; URL: http://dx.doi.org/10.1177/0093854811403123

[20] Gadher, D. (2020). 'Terrorism in the UK: Number of suspects tops 40,000 after MI5 rechecks its list'. *The Times. April 11*; URL: https://www.thetimes.co.uk/article/terrorism-in-the-uk-number-of-suspects-tops-40-000-after-mi5-rechecks-its-list-pqm6k62ph

[21] Victoroff, J. (2005). 'The Mind of the Terrorist: A Review and Critique of Psychological Approaches'. *Journal of Conflict Resolution*, *49*(1), 3-42; URL: https://doi.org/10.1177/0022002704272040