



Benchmarks for International Criminal Justice

**Feasibility Study Undertaken by the International
Nuremberg Principles Academy and HiIL**

Maurits Barendrecht, Isabella Banks (HiIL),
Juan Carlos Botero (Pontificia Universidad Javeriana),
and Jolana Makraiova (Nuremberg Academy)

April 2020



Preface

In 2019, the International Nuremberg Principles Academy (Nuremberg Academy) and The Hague Institute for Innovation of Law (HiIL) (“the project partners”) came together with Pontificia Universidad Javeriana in Bogotá, Colombia to launch a three-year research project exploring the feasibility of benchmarks for international criminal justice. It is with great pleasure that we present you the findings of this combined study today.

We concluded that establishing these benchmarks is not only feasible but, in our view, essential. A shared monitoring system can support goal-setting by helping to clarify the varied objectives of international criminal justice. It is also central to engaging with wider groups of stakeholders on the impact of the international criminal justice project, in furtherance of our shared mission to end impunity and achieve sustainable peace through justice. We explored the “how” and “why” questions and carefully considered the operational complexities of a global benchmarking system. Our key takeaways are described in the attached report and can be briefly summarised as follows:

1. Benchmarks for international criminal justice are needed.

The accumulated experience of benchmarks in other fields suggests that they can have a powerful transformative effect on critics of the international criminal justice project who question its return on investment. By enabling relevant stakeholders to identify and measure the contributions (or failings) of international justice mechanisms, benchmarks could facilitate a more data-driven assessment of their effectiveness and a more strategic allocation of funding to achieve the sector’s broader goals. Many monitoring systems are already in place in the justice sector, drawing from a range of data sources and facilitating priority-setting. Even when the data available is not comprehensive, quantitative measurement tends to spark problem-solving conversation and innovation.

2. Benchmarks for international criminal justice should measure the effectiveness of courts and transitional justice mechanisms alike.

Originally motivated by the legitimacy challenges faced by the International Criminal Court (ICC), the project partners quickly concluded that to be useful, the project scope must be broadened to include not only international criminal courts and tribunals (where internal procedural benchmarks are often already in place), but also transitional justice mechanisms. A wider approach was therefore adopted: one that took into consideration both the objectives arising from the Nuremberg legacy (including the [Nuremberg Principles](#)), and the broader goals of international peace and justice processes (like those carried out by the [Jurisdicción Especial para la Paz](#) (JEP)).

3. It is feasible to establish benchmarks for international criminal justice.

The study concluded that developing and maintaining such a monitoring system is substantively and operationally feasible, and that the benchmarks should be developed to measure the effectiveness of multiple actors in addressing international crimes.

1	Effectiveness of the response	<i>In achieving:</i> Prevention; deterrence; accountability; restoration; and other international criminal justice goals mentioned in Section 3 of this report.
2	By multiple actors	<i>Including but not limited to:</i> ICC ASP; ICC OTP; ICC Judges; ICC Registrar; ICC TFV; UNSC; UN General Assembly; UN IILMs; JEP and similar mechanisms; state organs; non-state actors, etc.
3	In addressing	<i>Handling the stages of the process from beginning to end in a timely, efficient, impartial, and effective manner:</i> Identifying; investigating; prosecuting; adjudicating; restoring, etc.
4	International crimes	<i>Including:</i> Situations in which international crimes such as genocide, crimes against humanity, war crimes, and the crime of aggression may have been committed.

Several foreseeable challenges were identified and include, inter alia, the need for strong partnerships and collaboration, an inclusive, objective, and methodologically rigorous approach, and excellent project management with long-term funding resources. The report recommends the following strategies for addressing these challenges and implementing the benchmarks in a sustainable way:

- The scope of the project should be limited enough to be feasible, but geographically and substantively inclusive enough to be broadly perceived as legitimate;
 - The audience for the benchmarks is wider than decision-makers in government and the international justice mechanisms themselves;
 - The project must be set up carefully to ensure robust project and risk management, methodological rigour and compliance with research ethics, long-term collaboration among stakeholders, and sustained funding;
 - The benchmarks themselves must be perceived as practical, applicable, objective, and legitimate;
 - A strong, independent partnership is needed to support and maintain the project. This includes preserving all data collected as part of the project and ensuring that it remains available for use by relevant stakeholders.
- 4. Developing benchmarks for international criminal justice is a resource-intensive exercise which will deliver the greatest returns in 5-10 years.**

Through desk research and a series of informal conversations, the project partners approximated the costs of the benchmarks and explored how they could be financed. From the information gathered, and by way of general indication only, it appears that the funding needed to set the project into motion in the first 3-4 years would be in the order of **6 million euros**. Running costs, once developed, would be in the order of magnitude of **1,5 to 2 million euros per year**. The project partners have assessed the most promising financing strategy to be **an initial investment of 3 million euros and a firm commitment of another 3 million euros**, with clear performance indicators and “go-no go” checkpoints throughout the 10-year development process.¹

¹ See Annex 4 for the detailed timeline and budget we estimate.

5. The Nuremberg Academy and HiiL remain committed to advancing this important conversation.

Benchmarks for international criminal justice would have a significant positive impact on the field. By helping international justice mechanisms articulate and organise around their common goals, assess their effectiveness, set priorities in an evidence-based way, and deliver more tangible and transparent results, such a system promises to offer more benefits than risks.

The Nuremberg Benchmarks proposed in the 2020 report are particularly timely now in light of increasing demands on the international criminal justice system. However, it is also clear that their establishment is a long-term undertaking. The project partners cannot implement this project as it exceeds their organisational mandates and operational capacities. A new coalition would need to be formed to carry this forward. To facilitate this next phase, the Nuremberg Academy and HiiL are sharing what they have learned in the process of assessing the project's feasibility. This knowledge has been summarised in the 2020 report that follows and disseminated in a side event at the 2021 Assembly of States Parties.

The project partners are grateful to the international experts who supported this research and welcome feedback on our findings. We remain committed to advancing the conversation around benchmarks for international criminal justice and highlighting the important contribution they can make to sustained peace.



Klaus Rackwitz
Director, International Nuremberg
Principles Academy
December 2022



Sam Müller
CEO, HiiL

December 2022

Table of Contents

Preface	1
Executive Summary	5
1. Background	6
2. Methodology	7
Core Questions	7
Selecting the Participants	7
Interviews	7
Two Workshops	8
The Workshop in Bogotá	8
The Workshop in The Hague	8
Records	8
3. Answers to the Core Questions	9
Foreseeable Benefits	9
Foreseeable Risks and Challenges	10
Design, Roll-out, and Tailoring	11
Scope	12
Target Users	13
Defining Success	14
Other Initiatives	15
Possible Criteria (Goals of International Criminal Justice)	16
4. Technical Dimensions of Measurement	17
Measuring Justice	17
The Relationship Between Research Question and Data Collection Methods	18
Key Dimensions of the Benchmark Development Process	19
5. Broad Strokes of Proposed Methodology	20
What is Measured	20
How it is Measured	21
Other Key Dimensions of Measurement	23
6. Limitations and Conclusion	25
Annexes	26
Annex 1: Added Value to Goals of Project Partners	26
Annex 2: Workshop Participants	27
Annex 3: Interview Participants	28
Annex 4: Estimated Timeline and Budget	29

Executive Summary

This report summarises a feasibility study that was conducted by the International Nuremberg Principles Academy (Nuremberg Academy) and the Hague Institute for Innovation of Law (HiIL), in collaboration with Pontificia Universidad Javeriana. The goal was to determine whether it would be feasible to establish benchmarks for international criminal justice, and if so, how these benchmarks could be created and under what circumstances.

Data was collected through two workshops with diverse groups of experts on international criminal justice and monitoring justice systems, one in Bogotá and one in The Hague. The authors relied on their extensive experience with monitoring rule of law, access to justice and effectiveness of justice interventions when sketching the elements of a possible monitoring system that addresses the challenges identified during the workshops.

The study concludes that developing and maintaining such a monitoring system is feasible. The accumulated experience of benchmarks in other fields, such as public health and climate change, has shown that the mere existence of data and benchmarks often has a powerful transformative effect. Even when data are not comprehensive, quantitative measurement tends to spark conversations and innovations. Benchmarks have helped to focus the conversation and set priorities in a number of different fields. Many monitoring systems now exist in the justice sector, using a broad range of data collection methods. The Nuremberg Benchmarks could become a system of indicators designed to measure the effectiveness of the response of multiple actors in addressing international crimes.

1	Effectiveness of the response	<i>In achieving:</i> Prevention; deterrence; accountability; restoration; and other international criminal justice goals mentioned in Section 3 of this report.
2	By multiple actors	<i>Including but not limited to:</i> ICC ASP; ICC OTP; ICC Judges; ICC Registrar; ICC TFV; UNSC; UN General Assembly; UN IIIMs; JEP and similar mechanisms; state organs; non-state actors, etc.
3	In addressing	<i>Handling the stages of the process from beginning to end in a timely, efficient, impartial, and effective manner:</i> Identifying; investigating; prosecuting; adjudicating; restoring, etc.
4	International crimes	<i>Including:</i> Situations in which international crimes such as genocide, crimes against humanity, war crimes, and the crime of aggression may have been committed.

The unit of analysis would be situations in which international crimes may have been committed: large scale atrocities in a country or region. The system would assess the effectiveness of the response in achieving a broad range of goals, including prevention, accountability, retribution or restoration. Actors would be the organisations that contribute to delivering international criminal justice through collecting evidence, arresting perpetrators, prosecuting, adjudicating and organising procedures. The way of addressing a situation includes the effectiveness of the response in handling stages of the process in a timely, efficient, impartial and cost-efficient manner.

1. Background

In 2018, representatives of the International Nuremberg Principles Academy and The Hague Institute for Innovation of Law (“the project partners”), and Pontificia Universidad Javeriana, met to discuss a potential project regarding benchmarks for international criminal justice.² This was based on the understanding that international criminal courts and similar accountability mechanisms should be working towards some indicators of success in order to increase their effectiveness and/or efficiency. The project partners agreed to explore whether a system of benchmarks could be developed in a feasibility study.

Starting around 1990, international criminal tribunals, the International Criminal Court (ICC) and a number of truth and reconciliation commissions have been set up. Since then, a wealth of experiences and data have been obtained which provide a firm foundation for the benchmark development process. Like many institutions, the ICC and other tribunals have been criticised for not meeting expectations, indicating a need to establish which expectations are realistic.

The Nuremberg Academy has published a series of studies regarding the deterrent effect of international criminal tribunals and the acceptance of these tribunals. In 2018, the Open Society Justice Initiative published a handbook intended to distil lessons learned from past mechanisms of criminal accountability for grave crimes and guide those designing new mechanisms. Moreover, advances in research and data collection methods have enabled new ways of evaluating and improving international criminal justice. It is now possible to inform stakeholders about outcomes and efficiency of their interventions. Compiling such information and bringing it into the public domain can help international justice stakeholders set priorities and improve the effectiveness of their efforts.

The goal of this study is thus to determine whether the project partners could develop a monitoring system that would help the stakeholders set strategic priorities in light of their responsibility to advance accountability for international crimes.

² The project partners thank Pontificia Universidad Javeriana for its support of this project.

2. Methodology

Core Questions

The project partners set out six questions that would guide the collection of data regarding feasibility:

- What are the foreseeable benefits (and risks) of creating these benchmarks?
- Which target groups would be the most instrumental to advance the project goals?
- How should these benchmarks be designed, rolled out, and tailored to key audiences, in order to enhance their practical usability and impact?
- Are there any other initiatives that are doing similar work, with which we should be consulting or collaborating before the project proceeds further?
- How should we define the success of these benchmarks?
- How far reaching should these benchmarks be (in terms of covering the ICC and other international criminal tribunals only, or ad-hoc, hybrid and national transitional justice mechanisms for mass atrocities - *e.g.* the Colombian Special Jurisdiction for Peace (JEP) - as well)?

To explore these questions, the project partners identified experts in international criminal justice, transitional justice, and social science research methods, and invited them to take part in one of two workshops. All those invited to participate were also asked to share their thoughts on the core questions in a brief individual interview prior to the workshops. The participant selection process, interviews, and workshops are described in detail below.

Selecting the Participants

The project partners looked for a diverse group of experts in terms of gender, age, regional representation, and disciplinary expertise. The final group included technical experts as well as leading scholars and practitioners in international criminal justice. In total, the project partners invited sixty experts to the workshops. Nineteen ultimately participated, and thirty-two experts in total were interviewed.

Interviews

The project partners requested individual interviews with the workshop invitees (regardless of their ability to attend their respective workshop) to clarify project goals and gather preliminary answers to the core questions. The answers provided in these interviews were analysed in preparation for the workshops and shared for further discussion during the workshops. This approach was taken to maximise the outcomes of the workshops, highlight recurring answers, and bring out fundamental disagreements.

In preparation for the Bogotá workshop, five participants were interviewed. In preparation for The Hague workshop, seventeen participants were interviewed. For a list detailing the institutional affiliations and areas of expertise of the ten interviewees who were unable to attend their respective workshop, please see Annex 3.

Two Workshops

Considering the project's focus on international criminal justice as a whole, the field of transitional justice, varying goals of international and internationalised tribunals, and specific situations such as Colombia, the project partners decided to organise two workshops to fully map out the project: one in Bogotá, Colombia, and the other in The Hague, Netherlands. The goal of the workshops was the same: to introduce, analyse, and answer the six core questions and advance discussions on the feasibility of the project. Participants in each workshop were also asked to discuss the goals of international criminal justice in order to inform the project scope.

For a list detailing the institutional affiliations and areas of expertise of the workshop participants, please see Annex 2.

The Workshop in Bogotá

The Bogotá workshop began by asking the participants to reflect on the six core questions. This approach was taken because the project partners were not able to interview all the Bogotá workshop participants in advance.

The afternoon session focused on identifying the goals of international criminal justice. The participants were asked to list the goals and organise them in terms of priority. The second afternoon session asked participants to identify the best way to measure and track the success of the Colombian peace process. The last exercise of the day asked participants to answer the same question with respect to Venezuela, Syria, Rwanda, Bosnia and Herzegovina, and Kenya.

The Workshop in The Hague

The workshop in The Hague followed a similar structure. It began by asking participants to brainstorm and explore the goals of international criminal justice. After this session, the participants discussed the interview results. These results were re-analyzed after the workshop in Bogotá and were therefore more comprehensive than those presented in Bogotá.

The afternoon session focused on risk assessment. The group was divided and assigned to one of three categories (strategic, quality/methodological and organisational/performance/sustainability) and asked to list all foreseeable risks under that category. The final exercise divided the group in half and asked them to explore two hypothetical scenarios: 1) What would have been different if there had been benchmarks in Bosnia and Herzegovina?; 2) What would have been different if there had been benchmarks in Kenya?

Records

The project partners produced summaries of the workshops and of the interviews with invitees/participants in the The Hague workshop. These records are available. The Bogotá interviews were more informal and primarily served as meetings to prepare for the workshops.

3. Answers to the Core Questions

Foreseeable Benefits

The workshop participants expressed a shared belief that developing benchmarks for international criminal justice could be useful in a number of ways. One of the primary benefits they foresaw was greater clarity around the goals and priorities of international criminal justice.

Commonly agreed upon goals could have secondary benefits for the field, such as shared expectations about what international criminal justice mechanisms can realistically achieve. Benchmarks could also incentivize improved performance by clarifying where responsibility lies for certain interventions and outcomes and contributing to a culture of evidence-based working. Participants agreed that all of these changes could help increase the transparency and legitimacy of international criminal justice mechanisms, which have come under intense scrutiny in recent years. The following table outlines the specific benefits the participants foresaw:

Both Workshops	The Hague Workshop
<p>Creating a common, cross-disciplinary understanding of the goals of international criminal justice. Goals vary depending on the mandates and targets set out by each mechanism. Clarifying what international criminal justice can and should achieve would help to bridge the expectations gap in the field.</p> <p>Assessing the effectiveness of international criminal justice institutions. Benchmarks could help shed light on the extent to which international criminal courts and tribunals have been successful in deterring international crime and closing the impunity gap. Such benchmarks would go beyond performance indicators like those being used at the ICC, which are just one small part of the bigger picture.</p> <p>Improving performance by putting pressure on international criminal justice actors. Benchmarks could achieve this by clarifying where responsibility lies for certain interventions and outcomes, while taking into consideration institutional capacities of the respective actors. This could create and align incentives and improve the overall performance of the international criminal justice field.</p> <p>Increasing the transparency and legitimacy of international criminal justice mechanisms. By setting out the roles, limitations, and mandates of each mechanism, benchmarks could clarify their relationship and interactions within the global justice system. Specifically, benchmarks could help mechanisms control the narrative about their work - a challenge which was raised in regard to the ICC as well as the JEP.</p>	<p>Facilitating priority-setting and decision-making by states by building a culture of evidence-based and comparative working. By contributing to a culture of evidence-based working, benchmarks could incentivize states to be better, more proactive stakeholders in the field. Increased availability of international criminal justice data would enable states to:</p> <ol style="list-style-type: none">1. Assess the justice needs of affected communities;2. Determine whether to invest in national or international mechanisms;3. Evaluate institutional performance;4. Learn from past mistakes and make more strategic decisions in the future;5. Better understand public perceptions of international criminal justice;6. Share knowledge;7. Balance competing priorities of peace and justice and build consensus about how best to address mass atrocities.

Foreseeable Risks and Challenges

The workshop participants identified numerous risks and challenges of benchmark development. These included strategic risks (such as lack of objectivity, misuse and abuse of the results, and the creation of perverse incentives); risks relating to benchmark quality and methodology (such as lack of credibility resulting from inadequate or biased measurement); and organisational risks related to performance and sustainability (such as failure to effect change in ways of working, and insufficient funding to sustain the project long-term). Participants agreed that while many of these risks could be mitigated, some (such as unintended consequences of developing benchmarks) would be uniquely challenging to anticipate and address. The following table outlines the specific risks and challenges the participants foresaw:

Both Workshops	The Hague Workshop
<p>Lack of credibility and acceptance. Benchmarks that are seen as unreliable or unacceptable (due to inadequate methodology or insufficient ownership, for example) will have little chance of success.</p> <p>Biased measurement and normative capture. Measurement choices may reflect the biases of the project partners, or the interests of the institutions the benchmarks are designed to measure. This could compromise the objectivity of the proposed project.</p> <p>Misuse and abuse. For example, the instrumentalization of benchmarks for budgetary control by states, or their weaponization by opponents of international criminal justice.</p> <p>Creating perverse incentives. Biased or inappropriate benchmarks could drive institutional purpose and practice in problematic ways.</p> <p>Raising false expectations. Benchmarks might contribute to unrealistically high expectations on the part of society - victims in particular - about what international criminal justice can achieve.</p> <p>Inappropriate, blanket application. While “global” benchmarks could facilitate comparisons across mechanisms, they could also be used to erase diversity in the field and enforce the hegemony of a particular approach.</p> <p>Conflict around what international criminal justice should achieve. International criminal justice goals and interests are often in tension. For example, the JEP’s focus on macro-criminality may be inconsistent with the case-by-case approach of international criminal courts and tribunals.</p> <p>Lack of follow-up. If no concrete action is taken following benchmark development, the proposed project will likely remain a theoretical exercise.</p> <p>Failure to produce practical, on-the-ground impact. The cost of developing the benchmarks may outweigh the practical value they add.</p>	<p>Lack of cooperation in obtaining data and evidence from the necessary stakeholders. International criminal justice institutions may be sensitive to criticism, particularly around issues they perceive as outside their mandate or beyond their control. Obtaining their cooperation, assistance, and data for the purpose of benchmark development may therefore prove difficult, unless they are engaged early in the process.</p> <p>Excessive bureaucratization. If international criminal justice mechanisms do choose to cooperate and share their data, there is a risk that they lose themselves in the process.</p> <p>Project proves unsustainable. The long-term funding necessary to sustain multi-year data collection and benchmark development may not be available. This risk increases with project scope, and would be particularly damaging to the quality of the benchmarks, due to the gap in periodic measurement it would create.</p> <p>Lack of diversity among project funders creates perceptions of bias. Because the proposed project is inherently political, the donors supporting it would need to be “the right donors.” Benchmarks developed with the support of five European countries, for example, would risk being perceived as biased.</p> <p>Culture shift toward evidence-based working does not take place. The success of the proposed project would depend on improved literacy in evidence-based working on the part of international criminal justice stakeholders. This kind of culture change presents a unique challenge: highlighting the lack of data-driven, user-centred decision-making among policymakers and justice professionals.</p> <p>Unintended consequences (the so-called, “unknown unknowns”). Unintended consequences resulting from fundamental issues with the benchmarks themselves would be among the most difficult risks to manage.</p>

Design, Roll-out, and Tailoring

The question of how benchmarks should be designed, rolled out, and tailored to key audiences in order to enhance their practical usability and impact was vigorously discussed in both workshops. Participants agreed that the benchmarks should be co-designed with clearly identified goals and measurement feasibility for the project in mind. They should be based on a mixed methods research and continuously adapted to better meet stakeholder needs. In discussing benchmark roll-out, participants emphasised the importance of accessibility and bottom-up ownership. All participants felt strongly that benchmarks should be tailored and specific to the mechanisms they were designed to measure.

The following table outlines the specific recommendations participants made with respect to benchmark design:

Both Workshops	Bogotá Workshop	The Hague Workshop
<p>Benchmarks should be co-designed with a diverse coalition of stakeholders. Project partners should consult stakeholders with a diverse range of perspectives, cultural backgrounds, and areas of expertise. An inclusive design process would help bring sceptics onboard and generate a sense of shared ownership of the result. At the same time, project partners should be conscious that diametrically opposed interests within the field make satisfying everyone an unrealistic goal.</p> <p>Benchmarks should be designed with clear goals in mind. Project partners should clearly establish what the benchmarks are intended to achieve. It may be useful to identify the primary gaps in international criminal justice as a starting point.</p> <p>Benchmarks should be designed with measurement feasibility in mind. Assisting with benchmark data collection must not be excessively burdensome for organisations on the ground. Benchmarks are only useful insofar as they are measurable.</p> <p>Benchmarks should be built on qualitative and quantitative measures. Whereas quantitative measures provide an indirect idea about quality, qualitative measures reflect values and are less easily misinterpreted.</p>	<p>Benchmarks should include goals and mechanisms from both international criminal justice and transitional justice paradigms. This could be achieved by focusing on shared goals of international criminal justice and transitional justice, such as prevention of violence and victim participation.</p> <p>Benchmarks should be designed around the needs of those most affected by international crime (victims).</p> <p>Benchmarks should measure every phase of the justice process, not just sentencing. Truth-telling and reintegration, for example, are fundamental elements of the Colombian peace process that would be missed by benchmarks that exclusively assessed sentencing outcomes.</p> <p>Designing positive (peace) benchmarks that relate reductions in violence to a particular justice mechanism may not be a good idea. This would be particularly difficult in Colombia, where the power vacuum left by the FARC has since been filled by other armed actors.</p>	<p>Benchmarks should be designed to measure the collective impact of the international criminal justice system as a whole, not just the performance of its parts. Benchmarks should account for the reality that achieving the goals of international criminal justice requires an interaction between many actors - each with a unique set of capacities, working conditions, and obligations. One way of achieving this would be to measure: 1) international organisations; 2) national justice institutions; 3) cross-country comparisons; 4) justice needs of individuals. This “four bucket” approach aims to measure and optimise the system as a whole, not only its individual parts.</p> <p>Benchmarks should measure popular perception of international criminal justice. This is challenging, but valuable. Benchmarks should recognize and take seriously the views of the majority while at the same time not sacrificing non-negotiable elements of international criminal justice.</p> <p>Benchmarks should focus on the goals and mechanisms of the international criminal justice paradigm. There is concern about the inclusion of wider transitional justice goals that were clearly outside the legal mandates of international criminal courts and tribunals.</p>

The following table outlines the specific recommendations participants made with respect to benchmark roll-out and tailoring:

Both Workshops	Bogotá Workshop	The Hague Workshop
<p>Project ownership should be shared across continents. The procedural legitimacy of the proposed project may be compromised if there is a perception that the benchmarks were imposed by Western Europe. Diverse perspectives are needed to deconstruct the dominant, eurocentric understanding of international criminal law.</p> <p>Benchmarks should be expressed in terms that everyone can understand. Benchmarks should be open, accessible, and accompanied by narratives that help people - particularly judicial actors - accurately interpret them.</p> <p>Benchmarks should be contextualised, tailored, and specific. Tailoring benchmarks to specific cases, situations, and mechanisms is important for reducing the risk of inappropriate application. Benchmarks should evaluate mechanisms on their own terms: according to their institutional mandates and the way they operate in practice.</p> <p>Benchmark design should be revisited and continuously improved upon. Benchmarks should be flexible, future-focused, and informed by lessons learned in the past.</p>	<p>Benchmarks that aim to promote transparency may be perceived as more authentic than those that aim to promote effectiveness. The latter may be viewed as propagandistic.</p>	<p>Benchmark development should start small and evolve gradually. Starting with a small, manageable project and setting up a pilot to test it conceptually would increase its chances of having an impact. At the same time, the scope of the benchmarks should not be so narrow that the benchmarks are reduced to institutional performance indicators and fail to capture key components of international criminal justice as a whole. Therefore, they must evolve over time.</p> <p>International criminal courts and tribunals should be actively engaged throughout the design process. Fostering a sense of institutional ownership and involving international criminal justice officials in the process without tying their hands is particularly important.</p> <p>The long-term nature of benchmark development should be clearly understood. This is important because benchmarks that are not updated with new data at regular intervals will fail to incentivize performance over time.</p>

Scope

Asked how far-reaching the benchmarks should be (in terms of covering the ICC and other international criminal tribunals only, or ad-hoc, hybrid, and national transitional justice mechanisms), the workshop participants clearly expressed that both the ICC and national systems prosecuting international crimes should be included. However, precisely defining the scope of the project proved difficult. Participants concluded that whereas broad benchmarks would be more difficult to develop and less likely to have an impact, narrow benchmarks would be less representative of the international criminal justice system as a whole and might therefore be regarded as less relevant. In addition to these broader considerations, participants in The Hague workshop suggested that the scope of the benchmarks could be informed by an assessment of the most pressing international criminal justice needs.

The following table outlines the specific recommendations participants made with respect to project scope:

Both Workshops	The Hague Workshop
<p>The ICC should be included. Inclusion of the ICC is important, both because of the legitimacy crisis it is currently facing and to ensure that the benchmarks are considered relevant.</p> <p>National prosecutions of international crimes should be included, but measured differently than international justice mechanisms. Whereas benchmarks for international justice mechanisms should be developed through consensus of the international community, benchmarks for national prosecutions of international crimes should be dynamic and context-specific. Making comparisons across national and international justice systems risks lowering the standard baseline of the benchmarks. Defining the project scope based on this baseline or “lowest common denominator” may help to ameliorate this risk.</p> <p>Narrowing the scope of the benchmarks may increase their impact. Narrowing the benchmarks to the institutional framework of international criminal justice would likely make them more realistic, understandable, impactful, and sustainable.</p> <p>If the scope is too broad, the benchmarks will not be coherent, but if the scope is too narrow, the benchmarks will not be relevant. A broader approach would allow diverse mechanisms to be compared and shared challenges identified. On the other hand, a “one-size fits all” approach may be ineffective. In light of this tension, it was agreed that benchmarks should focus on national and international institutions, at least in the near-term.</p>	<p>If the benchmarks are developed to apply to the international criminal justice system as a whole, it is important to measure factors outside of the control of international criminal courts and tribunals. While the courts would likely prefer benchmarks that exclusively measured factors within their control, limiting the project scope in this way would not benefit the international criminal justice system as a whole.</p> <p>One way of strategically narrowing the scope could be to begin benchmark development with an assessment of the most pressing justice needs. Taking a user-centred, needs-based approach could increase the usefulness of the benchmarks and incentivize international criminal justice institutions to justify themselves on this basis.</p> <p>Even if the scope was limited to international criminal courts and tribunals, the benchmarks should cater to the specifics of each mechanism type. Benchmarks could, for example, distinguish between the ICC, ad-hoc tribunals, hybrid tribunals, and the independent, impartial, and international mechanisms (IIIMs).</p>

Target Users

The concept note shared with workshop participants outlined the expected target groups for the proposed project. These included decision-makers - especially state representatives and policymakers who are involved in setting out state policy priority lists - as well as NGOs and other organisations promoting effective international criminal justice. The workshop participants generally agreed with the proposed target groups, but struggled with the question of whether it was more important for the benchmarks to target experts or the general public.

The two dominant perspectives shared by the participants with respect to target users were:

Both Workshops

Benchmarks should primarily target influential decision-makers. Because benchmarks are an instrument, they should target those who can influence policy. These target users include international justice institutions and their representative bodies, states and their representatives, politicians, NGOs, and the media. While the public may be a beneficiary of the benchmarks, it is not accurate to classify it as the target user.

Benchmarks should primarily target the general public. Influential actors like justice leaders, states, politicians, NGOs, and the media get their strength and influence from the public. Therefore, benchmarks should cater to their needs. Victims in particular should be prioritised, given that they are the most harmed by international crime and the most often neglected.

Despite this difference of opinion, the majority of participants in The Hague workshop seemed to agree with the comment that, “The ultimate answer might be that some benchmarks are for specific target users, and others are for the public.”

Defining Success

Visions of success shared by the participants centred around benchmark clarity, dynamism, acceptance, practical impact, and connection to the public. Participants in both workshops stressed that in addition to affecting change in the field, benchmarks should, in the long-run, carry broader societal meaning and contribute to the public’s understanding of justice. It is important to note that while the visions of success provided below are not mutually exclusive in theory, they cannot realistically be achieved at once. The following list is intended to faithfully reflect what was discussed in the workshops, and should not be understood as a list of practical goals for the Nuremberg Benchmarks.

The workshop participants felt that the benchmarks would be successful insofar as they:

- Made clear from the start what they set out to achieve.
- Established an accurate and objective system of measurement.
- Were balanced (in terms of priorities) and evolved gradually.
- Were accepted, referenced, and acted upon by organisations in the field.
- Stimulated judicial institutions to improve their performance.
- Resulted in more efficient spending on international criminal justice mechanisms.
- Were accepted as credible across disciplines.
- Were general, understandable, and communicated in such a way that all parts of society could have a relationship with them.
- Contributed to public understanding of the meaning of justice.
- Helped to reduce false expectations about international criminal justice.
- Contributed to more satisfying processes and outcomes for victims.
- Contributed to stability, non-repetition, and positive peace.
- Helped to align peace and justice processes.
- Helped to reduce the impunity gap.

- Were internationally recognized and used by others.
- Were practical and adaptable.
- Were dynamic and continuously improved upon.

Other Initiatives

During the pre-workshop interviews, participants noted that there may be opportunities for synergies between the proposed project and other similar initiatives. Some of these benchmark projects focus on measuring court performance, whereas others look more broadly at the legitimacy and net value of international criminal justice to society.

The information provided is categorised thematically in the following table:

Measuring the Performance of International Criminal Courts and Tribunals	Measuring the Legitimacy of International Criminal Courts and Tribunals:	Measuring the Value of International Criminal Justice
<p>Institutions</p> <p>ICTY</p> <p>ICC</p> <p>Leiden University</p> <p>Monash University</p>	<p>Institutions</p> <p>Temple University</p>	<p>Institutions</p> <p>University of Toronto</p>
<p>Specific Initiatives</p> <p>Strategic Plan of the ICC Office of the Prosecutor</p> <p>Paris Declaration on the Effectiveness of International Criminal Justice</p> <p>Dakar Guidelines on the Establishment of Hybrid Courts</p>		

Possible Criteria (Goals of International Criminal Justice)

In addition to addressing the six core questions presented in the concept note, workshop participants discussed what specifically the benchmarks should measure. In attempting to clarify the goals of international criminal justice, participants in both workshops struggled to decide whether the goals should be realistic and limited, or aspirational yet potentially misleading and frustrating for victims. Whereas in Bogotá this remained more of an open question, participants in The Hague workshop resolved this perceived tension by drawing a distinction between what they saw as the primary goals of international criminal justice (*i.e.* accountability) and its secondary effects (*i.e.* prevention). Though reaching any consensus proved difficult, participants ranked the goals (roughly - see notes) as follows. Colour coding is used to highlight similar goals across categories.

International Criminal Justice (The Hague Workshop)*	International Criminal Justice (Bogotá Workshop)**	Colombian Peace Process (Bogotá Workshop)**
<p>1. Effective and efficient international courts and organisations (10 votes)</p> <p>2. Strengthening/supporting national justice systems (8 votes)</p> <p>3. Strengthening the global justice system (6 votes)</p> <p>4. Contribution to peace and stability (6 votes)</p> <p>5. Victims' redress (6 votes)</p> <p>6. Strengthening the rule of law and human rights (5 votes)</p> <p>7. Ending impunity (5 votes)</p> <p>8. Prevention (2 votes)</p> <p>9. Truth (2 votes)</p> <p>10. Public condemnation/affirming community norms (1 vote)</p> <p>11. Being an effective last resort (1 vote)</p>	<p>1. Prevention of international crime</p> <p>2. Fighting impunity/accountability</p> <p>3. Victim satisfaction</p> <p>4. Strengthening domestic proceedings</p> <p>5. Due process/procedural fairness</p> <p>6. Contribution to truth and historical record</p> <p>7. Promotion of human rights and rule of law values</p> <p>8. Restoration</p> <p>9. Reparations</p> <p>10. Retribution</p> <p>11. Deterrence</p>	<p>1. Prevention of violence</p> <p>2. Quality of case prioritisation and selection</p> <p>3. Contribution to truth</p> <p>4. Offender reintegration</p> <p>5. Victim participation</p> <p>6. Extent to which justice is seen to be done</p> <p>7. Quality of implementation</p>

* Participants in The Hague workshop collectively brainstormed the primary goals of international criminal justice and then voted on which among these they considered to be the five most important.

** Because these goals were weighted in small groups rather than collectively ranked by all participants in the Bogotá workshop, their order reflects estimates made by the workshop organisers about what the group considered most important.

4. Technical Dimensions of Measurement

One of the primary aims of the feasibility study was to determine whether it was possible to develop a methodologically rigorous and sustainable system of benchmarks to measure international criminal justice. Whereas the previous sections outlined the implications of such a system for international criminal justice stakeholders and society at large, the sections that follow introduce its technical dimensions. Specifically, they identify key methodological questions that must be answered in order to narrow the scope of the proposed project and construct the Nuremberg Benchmarks within a reasonable period of time.

Measuring Justice

Justice can be measured in many different ways. Two well-known examples illustrate the point:

1. **Measuring justice as an institutional process.** In 2002 the Council of Europe created the European Commission for the Efficiency of Justice (CEPEJ), which is composed of representatives of 47 member states.³ CEPEJ's benchmarks collect, analyse and publish comparative data on justice's inputs, outputs, and outcomes, from the perspective of the justice institutions themselves. These are mainly *supply-driven* benchmarks, which use judicial statistics as the main source of data.
2. **Measuring justice as a user-centred service.** At the other end of the spectrum, for over a decade The Hague Institute for Innovation of Law has been collecting, analysing and publishing comparative data on justice's inputs, outputs, and outcomes, from the perspective of the system's users. These are mainly *demand-driven* benchmarks, which use justice needs surveys as the main source of data.

Both systems of indicators largely aim at measuring the same thing: the comparative efficiency, effectiveness, fairness, and impartiality of justice delivery across countries. However, they use very different data collection methods. The choice of methods in both cases is not simply methodological—it is driven by different 'research questions' or perspectives.

³ "The aim of the CEPEJ is the improvement of the efficiency and functioning of justice in the member States, and the development of the implementation of the instruments adopted by the Council of Europe... In order to carry out these different tasks, the CEPEJ prepares benchmarks, collects and analyses data, defines instruments of measure and means of evaluation, adopts documents (reports, advices, guidelines, action plans, etc), develops contacts with qualified personalities, non-governmental organisations, research institutes and information centres, organises hearings, promotes networks of legal professionals." CEPEJ's web page: <https://www.coe.int/en/web/cepej/about-cepej> Access on Oct. 7, 2019.

The Relationship Between Research Question and Data Collection Methods

A time-tested principle in social science research postulates that *method follows question, i.e.*, researchers must first define the research question that they are trying to answer, and only then select the methodology that is most suitable to answer it.⁴ A first corollary of this principle is that not all data collection methods are suitable to answer all research questions. The same logic applies to the process of developing “benchmarks” or “indicators.” De Vaus explains this point clearly⁵:

“An analogy might help. When constructing a building there is no point ordering materials or setting critical dates for completion of project stages until we know what sort of building is being constructed. The first decision is whether we need a high rise office building, a factory for manufacturing machinery, a school, a residential home or an apartment block. Until this is done we cannot sketch a plan, obtain permits, work out a work schedule or order materials.

Similarly, social research needs a design or a structure before data collection or analysis can commence. A research design is not just a work plan. A work plan details what has to be done to complete the project but the work plan will flow from the project's research design. The function of a research design is to ensure that the evidence obtained enables us to answer the initial question as unambiguously as possible. Obtaining relevant evidence entails specifying the type of evidence needed to answer the research question, to test a theory, to evaluate a programme or to accurately describe some phenomenon. In other words, when designing research we need to ask: given this research question (or theory), what type of evidence is needed to answer the question (or test the theory) in a convincing way?”

As it follows from the previous section, workshop participants expressed different expectations about what a system of benchmarks of international criminal justice could and should be. In technical terms, when participants were asked about potential *benefits* and *risks* associated with these benchmarks, they had different ‘research questions’ in mind. Importantly, these diverging views and expectations among workshop participants are probably indicative of the wide range of views and expectations of the broader community of stakeholders for this project.

This divergence of expectations creates a methodological “*chicken and egg*” problem for this project between the *what* (research question) and the *how* (data collection methods). Given limited resources, it is unrealistic to expect that all data collection methods that would be necessary to address the multiple expectations (or research questions) of the various stakeholders, may be employed.

If unlimited resources were available for this project, theoretically it would be possible to develop a system of indicators so comprehensive that all (or nearly all) possible expectations were taken into account, and proxy measures were developed to answer multiple research questions. This does not seem to be the case. From the outset, the project partners were mindful about the need to focus and prioritise the use of resources, to ensure the benchmarks’ long-term sustainability. As the workshop outcomes make clear, sustainability is of the essence—since the ultimate goal of benchmarks is to identify problems in order to drive improvement, most potential gains from benchmarks come from periodic measurement, which enables tracking of progress (or lack thereof) across time.

⁴ See, e.g., Neuman, W. L. *Social research methods: Qualitative and quantitative approach* (6th ed.) Upper Saddle River: Pearson, 2006.

⁵ De Vaus, David. *Research Design in Social Research*. Sage Publications. 2001. 2006 Reprint. Pp. 8-9.

Choices must be made in order to narrow the scope of expectations about exactly what—and at what level of precision and detail—may reasonably be measured by a set of benchmarks for international criminal justice that aspires to periodic data collection and sustainability across time. While it will not be possible to satisfy the widely diverging expectations among ICJ stakeholders, the work carried out by the project partners during the past 16 months, and contributions from workshop participants in Bogotá and The Hague, have enabled us to sketch some broad strokes of a methodology that could simultaneously take into account multiple goals, actors, and expectations for these benchmarks, while ensuring methodological rigour and sustainability of the project. The pages that follow outline the key dimensions that must be taken into account to develop these benchmarks.

Key Dimensions of the Benchmark Development Process

All systems of indicators have five key dimensions⁶:

1. **What** is measured. This encompasses a wide range of issues, including among others, *Subject matter, Perspective, Unit of analysis; and Selection of variables.*
2. **How** it is measured. This dimension refers to *Data collection sources and methods; Imputation of missing data, normalisation, aggregation and weighting; and Data analysis and robustness checks,* among other issues.
3. **Why** it is measured. This refers mostly to *intended targets and Intended effects.*
4. **How findings are presented.** There are choices to be made about the way findings are presented. For instance: *Comparative vs. case-specific; Numeric vs. narrative; Precise rankings vs. wider bands.*
5. **Mitigation of unintended effects.**

These dimensions are expanded upon in the next section, taking into account feedback and suggestions from workshop participants and expert interviews.

⁶ While these five dimensions are the most important, there are a number of additional features to consider while creating a system of indicators in the broad policy area of justice and the rule of law. See, e.g. OECD & EU-JRC.2018. *Handbook on Constructing Composite Indicators*; Botero, JC, Nelson, R, and Pratt, C. "Indices and indicators of justice, governance and the rule of law: an overview" *Hague Journal on the Rule of Law*, Vol. 3, Issue 2 (2011), pp. 153-169; Botero, JC, Pinzon, AM, and Pratt, C. "How, when and why do governance, justice and rule of law indicators fail public policy decision making in practice?" *Hague Journal on the Rule of Law*, pp 1-24. Springer International Publishing, January 2016.

5. Broad Strokes of Proposed Methodology

The previous section established that in order to construct the Nuremberg Benchmarks, it is necessary to specify **what** exactly the benchmarks would measure (the research question), and **how** (the method). This section sets out a framework for doing so, based on the expert opinions shared during and prior to the workshops. It also discusses other key dimensions of benchmark development—including intended targets and effects (the **why**), presentation of findings, and mitigation of unintended consequences—all of which can have significant implications for the overall impact of the proposed benchmarks.

What is Measured

During the course of the workshops, it became apparent that for this project to succeed, the Nuremberg Benchmarks would need to be able to measure the performance of *multiple actors* in relation to *multiple ICJ goals*, through *various stages of the process*, and from *multiple perspectives*. A tentative formula for this complex endeavor emerged. This formula reads as follows:

*“The Nuremberg Benchmarks are a system of indicators designed to measure the **effectiveness of the response by multiple actors in addressing international crimes.**”*

This formula may be disaggregated in multiple ways. The following formulation is offered only by way of example:

- **Effectiveness of the response** in achieving Prevention, Deterrence, Accountability, Restoration, and other international criminal justice **goals** mentioned in Section 3 of this report.
- By **multiple actors**, including the ICC Assembly of States Parties; ICC OTP; ICC Judges; ICC Victims Fund; ICC Registrar; UN Security Council; UN General Assembly; UNHRC; ICTY OTP, Judges, Registrar, etc.; ICTR OTP, Judges, Registrar, etc.; UN IM on Syria; Colombian Special Jurisdiction for Peace, and similar mechanisms; Various organs of the state where crimes were perpetrated; Non-state actors; Etc.
- In **addressing**, *i.e.*, handling the stages of the process from beginning to end, in a timely, efficient, impartial and effective manner: Identifying; Investigating; Prosecuting; Adjudicating; Restoring; etc.
- **International crimes**: Genocide; Crimes against humanity; War crimes; Crime of aggression; Other?

Under this tentative framework, the Nuremberg Benchmarks would measure the effectiveness of the actions taken (or not taken) by each relevant ICJ actor, with respect to various goals, at different stages of the process.

The **unit of analysis** of this measurement exercise would be the “situation,” meaning the situation present at a certain place during a certain period of time (*e.g.* the Syrian situation or the Libyan situation, etc.) The determining factor for measurement purposes is not whether the ICC has jurisdiction or whether some authority has defined or framed the situation in a certain way, but rather whether the information available provides a reasonable basis to believe that international crimes may be present in a particular country or territory at a certain point in time. One example may help to illustrate the idea:

Example: Actions taken (and not taken) by the ICC OTP with respect to the Colombian situation in 2018.

According to the ICC-OTP's *Report on Preliminary Examination Activities*, of December 5, 2018 (p. 44),⁷ the OTP's actions (or inaction) with respect to the Colombian Situation, are as follows:

"Conclusion and Next Steps

164. In the context of its ongoing admissibility assessment, the Office will continue to engage with the Colombian authorities to receive information on concrete and progressive investigative steps and prosecutorial activities carried out with respect to the potential cases it has identified. By the same token, the Office will continue engaging with other stakeholders who continue to inform the OTP's assessment of the situation.

165. The OTP will continue assessing the genuineness of the proceedings carried out under the ordinary justice system, the JPL tribunals and the SJP. While noting with appreciation that the SJP is now fully operational, the OTP will continue examining developments relating to its regulations, operations and proceedings to the extent that the functioning of the jurisdiction will have a critical impact on the OTP's assessment of the admissibility of potential cases arising out of the situation in Colombia. In this context, the OTP will closely follow individual proceedings that arise from the cases initiated so far, as well as the identification of new cases selected for investigation and prosecution."

This action, *i.e.*, to "continue assessing the genuineness of the proceedings..." can be assessed with respect to various ICJ goals. For the sake of argument, the same action may be assessed today as highly positive from the perspective of strengthening local capacity, while simultaneously being assessed as somewhat positive from the perspective of accountability or victims' redress. If the Colombian situation continues unaltered for the next year, and ICC-OTP's actions remain the same, these actions could then be assessed somewhat differently a year from now. As in the case of Kenya, the absence of direct intervention by the OTP may be assessed in different ways at different times.

For measurement purposes, what is essential is to have a clear definition of **what** it is that the benchmarks are measuring. In the Colombian example, this particular benchmark would measure the actions of one actor (the ICC-OTP), with respect to multiple goals. Other benchmarks would measure the actions of other actors (*e.g.*, the Colombia SJP, or the UNSC), with respect to various ICJ goals. The net result of the whole measurement exercise would be a complex matrix of hundreds of variables that measure a handful of goals with respect to a handful of actors, regarding a handful of situations, at different stages of the process, from a handful of different perspectives. All these variables would then be measured, normalised, and indexed in the way described in the next section.

How it is Measured

This dimension refers to *data collection sources and methods; imputation of missing data, normalisation, aggregation and weighting; and data analysis and robustness checks*, among other issues.

As mentioned above, *method follows question, i.e.*, researchers must first define the research question that they are trying to answer, and only then select the methodology that is most suitable to answer it. If this project goes forward, some of the first steps would involve the design of a methodology to narrow the research question (*i.e.*, the

⁷ International Criminal Court, 2018. [Report on preliminary examination activities](#).

what is measured). This methodology would aim at identifying a manageable list of ICJ goals, for a reasonable list of ICJ actors, with respect to a manageable list of ICJ situations, from a handful of perspectives.

While this is a complex and time-consuming task, the technologies to do it are readily available. Several workshop participants suggested that the process and methodology employed by the World Justice Project to develop the *WJP Rule of Law Index* (described in detail in Botero & Ponce, "Measuring the Rule of Law"),⁸ would provide useful guidelines for this stage of the Nuremberg Benchmarks project. As in the case of the *WJP Index*, this process would include a series of steps, consultations, pilot-testing, and staggered deployment of highly tailored measurement instruments. The overall goal of this first stage of the process would be to narrow the project's research question, as it was explained above.

A second (concurrent) step of the measurement process would entail the selection of data sources and methods. Since *method follows question*, it is not possible to define at this early stage possible data sources and data collection methods with precision. Based on the work of the past year, and feedback from the workshops in Bogotá and The Hague, the project partners tentatively propose the following conclusions:

1. Research must be multi-method, including both quantitative and qualitative data collection methods.
2. Multiple sources of data would be used, including:
 - a. The bulk of information would come from relatively inexpensive expert questionnaires. This would start with extensive consultation with ICJ experts, initially taking the form of semi-structured questionnaires and interviews, and ultimately leading to structured instruments. The end product would be a standardised questionnaire, which may not only be tailored to particular ICJ "situations," but may also be deployed on a yearly basis at a relatively low cost, to ensure the project's long-term sustainability. Multiple methodologies are readily available to develop these instruments. The *WJP Rule of Law Index* Qualified Respondent Questionnaire, and Freedom House's *Freedom in the World* questionnaire,⁹ may be examples of the kind of standardised questionnaire that would ultimately emerge from this process.
 - b. Limited probability sample polling may be necessary,¹⁰ together with extensive non-probability sampling of the general public or specific populations, e.g., Rohingya refugees in Bangladesh, or Syrian refugees in Turkey.
 - c. Rigorous qualitative fieldwork (mainly focus groups) would complement the picture.
 - d. Data from the ICC and other tribunals and mechanisms (such as the Colombian SJP), would also be collected. These data would include both process and outcome indicators.
3. Triangulation among data sources and methods would be necessary, both to ensure data reliability and in order to address the problem of incentives of institutional indicators. This is described in detail in Botero *et. al*, "Indices and Indicators of Justice, Governance and the Rule of Law: An Overview," at pages 161-163.¹¹

⁸ Botero, Juan Carlos and Ponce, Alejandro, *Measuring the Rule of Law* (November 30, 2011). Available at SSRN: <https://ssrn.com/abstract=1966257> or <http://dx.doi.org/10.2139/ssrn.1966257>

⁹ Abramowitz, M.J. *Freedom in the world 2018. Democracy in crisis*. Available at <https://freedomhouse.org/report/freedom-world/freedom-world-2018> Accessed on September 15, 2023.

¹⁰ A state of the art on legal needs surveys was recently published by the OECD. See OECD/ Open Society Foundations (2019), *Legal needs surveys and access to justice*, OECD Publishing, Paris.

¹¹ Botero, JC, Nelson, R, and Pratt, C. "Indices and indicators of justice, governance and the rule of law: an overview" *Hague Journal on the Rule of Law*, Vol. 3, Issue 2 (2011), pp. 153-169.

Other methodological aspects of the **how** question include *imputation of missing data; normalisation, aggregation and weighting; and data analysis and robustness checks*. Several methodologies and examples are readily available to do this work. The description of these steps for the *WJP Rule of Law Index* in Botero & Ponce,¹² may be used as a specific example of the kind of steps involved. More generally, the OECD & EU-JRC *Handbook on Constructing Composite Indicators*,¹³ contains a comprehensive list of alternative methodologies available.

Other Key Dimensions of Measurement

Finally, a few words on the last three dimensions of the benchmark development process (namely the why question; the choices on presentation of results; and the measures employed to mitigate unintended effects). These dimensions have been extensively discussed in other segments of this report.

The **why it is measured** question refers mostly to *intended targets* and *intended effects* of the indicators. Considerable thought must be given to assess in advance not only the specific target of each benchmark, but also how and to what end it is targeted. Ultimately, this involves a ‘theory of change’ consideration on the incentives that indicators produce on their intended targets, *i.e.*, how different benchmark choices may produce different incentives.

Secondly, there are choices to be made about the way findings are presented that have significant consequences on the overall effectiveness and impact of the benchmarks. These choices include, for instance: *comparative vs. case-specific; numeric vs. narrative; precise rankings vs. wider bands*. Botero, Nelson and Pratt,¹⁴ offer some recommendations in this respect—a number of which were raised in the expert workshops:

“Several strategies have been designed by indicator developers to prevent these abuses. First, to release disaggregated data; instead of providing a single number per country, it is generally more useful to provide scores for a variety of dimensions covered by the indicators. Second, to offer detailed qualitative descriptions of the findings. Third, to provide as much context as possible, including details about the conceptualization of the indicators and the data collection and aggregation methods employed to produce them. Fourth, to the extent possible, to avoid cross-country comparisons altogether. Fifth, if cross-country comparisons are made, to encourage ‘apple to apple’ comparisons, i.e., comparisons among similarly situated countries (by region, income-level or other relevant characteristics that may help users of indicators to place the findings in context). Sixth, to design the indicators in a manner that would make it more difficult to target them; for instance, a single indicator about regulatory performance is more vulnerable to government attacks than a composite score encompassing a variety of dimensions of regulatory effectiveness. Seventh, to the extent possible, to avoid data collection methods that may be subject to government control, such as official statistics and self-reported data by government officers or interested parties. Eighth, if government data are collected, to test the convergent validity of the data.”

Finally, all benchmarks have unintended consequences. Specific examples of unintended consequences of indicators developed by governments and other actors, which produced catastrophic effects (such as the Colombian “false positives” case that is currently under preliminary examination at the ICC OTP), are described in detail at

¹² Botero, Juan Carlos and Ponce, Alejandro, *Measuring the Rule of Law* (November 30, 2011). Available at SSRN: <https://ssrn.com/abstract=1966257> or <http://dx.doi.org/10.2139/ssrn.1966257>.

¹³ OECD & EU-JRC (2008) *Handbook on Constructing Composite Indicators*.

¹⁴ Botero, JC, Nelson, R, and Pratt, C. “*Indices and indicators of justice, governance and the rule of law: an overview*” *Hague Journal on the Rule of Law*, Vol. 3, Issue 2 (2011), p. 159.

Botero, Pinzon & Pratt.¹⁵ Thus, it is necessary to adopt strategies to mitigate these unintended effects.

There are several key requirements that must be met by indicators - all types of indicators, from those developed by local government agencies based on official statistics, to those produced by international organisations for cross country analysis - in order to be both technically acceptable and of practical utility. First, the quality of the conceptualisation of what is being measured is extremely important, and it's crucial for end-users to understand the underlying assumptions and value structure of what is being measured. Second, one must check the indicators' technical dimensions, such as the rigor of the data collection, aggregation, imputation, weighting, and normalisation methods which are used to produce them. In addition, uncertainty and sensitivity analyses, and other methods of explicit reporting of margins of error, are essential tools to understand the meaning of numbers. While these statistical analyses are generally beyond the reach of ordinary citizens, they cannot be ignored by governments, donor agencies and other constituencies who attempt to base or track policy decisions on these indicators. Finally, specific indicators must be used in context.

An effective indicator system not only provides information on whether and to what extent progress is being made in one particular aspect, but also how progress in achieving one government objective may negatively affect another. This is true at the micro level, such as the case of a local police chief trying to solve more crimes with less intrusion on citizens' liberties - which is clearly described by Foglesong and Stone in another piece of this publication. It is also true at the macro level, such as the uneasy interaction between guaranteeing order and security at the country level, while providing effective protection of fundamental rights in low and middle-income countries, as suggested by the WJP Rule of Law Index data. An effective system of indicators tracks different dimensions of the system together, and provides information about interaction among these dimensions over time. This is particularly important after intervention measures are implemented.¹⁶

The importance of such mitigation strategies was made clear during workshop discussions around benchmark design and risks, which emphasised the need to measure the interaction of various international actors and the net societal impact of the international criminal justice system.

¹⁵ Botero, JC, Pinzon, AM, and Pratt, C. "How, when and why do governance, justice and rule of law indicators fail public policy decision making in practice?" *Hague Journal on the Rule of Law*, pp 1-24. Springer International Publishing, January 2016.

¹⁶ Botero J, Martinez J, Ponce A, Pratt C (2012) The rule of law measurement revolution: complementarity between official statistics, qualitative assessments, and quantitative indicators of the rule of law. In: Botero et al (ed) *Innovations in rule of law—a compilation of concise essays*. HiiL and the World Justice Project. <https://www.hiil.org/research/innovations-in-rule-of-law/>, pp. 9-10.

6. Limitations and Conclusion

The goal of this study was to determine whether it could be feasible to establish benchmarks for international criminal justice, and if so, how these benchmarks would be created and under what circumstances. Our conclusion is that developing and maintaining such a monitoring system is feasible. The expected benefits, as seen by experts, are considerable and are likely to outweigh the costs. Risks have been identified, but can be managed, building on the experience collected when operating the many monitoring systems that now exist in the field of justice, using a broad range of data-collection methods, and proceeding in a methodical, step-by-step way.

The limitations of this feasibility study have to be taken into account. The experience of experts consulted during the workshops and the interviews was impressive, but their number was limited. By working in two locations (The Hague and Bogotá), we may have missed other perspectives. We shared the report with experts for additional feedback, including section 5, in May 2020, seeking additional feedback. We tested this model ourselves against the risks, benefits and other elements that determine feasibility, including the project budget and complexity.

Participant feedback on this feasibility study is welcome and should be directed to the contacts provided on the final page of this report.

Annexes

Annex 1: Added Value to Goals of Project Partners

The **International Nuremberg Principles Academy (Nuremberg Academy)** is a non-profit foundation dedicated to the advancement of international criminal law and related human rights. It was established by the Federal Republic of Germany, the Free State of Bavaria, and the City of Nuremberg in 2014. The Nuremberg Academy is located in Nuremberg, the birthplace of modern international criminal law, where the Nuremberg trial against the major war criminals was held by the International Military Tribunal from 1945 to 1949. For the first time in history, an international tribunal was authorised to hold leading representatives of a state personally accountable for crimes under international law. The foundation carries forward the legacy of the Nuremberg trials and the “Nuremberg Principles”, which comprise the principles of international law recognized in the Charter of the Nuremberg Tribunal and in the judgement of the Tribunal. The International Law Commission, a body of experts established to help develop and codify international law since 1947, was entrusted to formulate these principles, which was finalised in 1950. Conscious of this historic heritage, the Nuremberg Academy supports the fight against impunity for universally recognised international core crimes: genocide, crimes against humanity, war crimes, and the crime of aggression. Its main fields of activity include providing a forum for dialogue by convening conferences and expert meetings, conducting interdisciplinary and applied research, engaging in specialised capacity building for practitioners of international criminal law, and human rights education. Dedicated to supporting the worldwide enforcement of international criminal law, the Nuremberg Academy upholds the Nuremberg Principles and the rule of law with a vision of sustainable peace through justice, furthering knowledge, and building capacities of those involved in the judicial process in relation to these crimes.

HiiL (The Hague Institute for Innovation of Law) is a social enterprise devoted to user-friendly justice. That means justice that is easy to access, easy to understand, and effective. We will ensure that by 2030, 150 million people will be able to prevent or resolve their most pressing justice problems. We do this by stimulating innovation and scaling what works best. We are friendly rebels focused on concrete improvements in the lives of people. Data and evidence are central in all that we do. The Nuremberg Benchmarks project is closely aligned with HiiL’s mission to ensure that justice is easy to access, easy to understand, and effective. As part of the UN Sustainable Development Agenda, HiiL is committed to supporting the transition towards a more people-centred justice sector - one that prevents justice problems and helps people find solutions - in order to achieve equal access to justice for all. In the field of international criminal justice, this means using data and evidence to identify gaps in current ways of working and increasing pressure on global actors that are not doing their part to close the impunity gap. HiiL believes that developing benchmarks in partnership with a diverse coalition of global stakeholders is a promising way of measuring international criminal justice and achieving its broader goals.

Annex 2: Workshop Participants

Bogotá	
Institutional Affiliation	Area of Expertise
U.N. University for Peace	International law, human rights law and international relations
Office of the Federal Prosecutor, Republic of Mexico	International criminal law and transitional justice
Pontificia Universidad Javeriana Law School	Domestic criminal justice, international human rights, and empirical legal research methods
Pontificia Universidad Javeriana Law School	International criminal law and the Special Jurisdiction for Peace (JEP)
Universidad de Los Andes	Criminal law and the International Criminal Court
Universidad de la Sabana	International law and the legitimacy of international institutions
International Center for Transitional Justice	Transitional justice, punishment in liberal societies, restorative justice
Ombudsman's Office	International criminal justice and the Special Jurisdiction for Peace (JEP)

The Hague	
Institutional Affiliation	Area of Expertise
World Bank	Justice reform initiatives and justice sector analysis
Special Tribunal for Lebanon	Criminal justice, transitional justice, rule of law, and human and women's rights
Dutch Ministry of Foreign Affairs	International criminal law and transitional justice
Dutch Ministry of Justice and Security	Legal and policy issues of international criminal justice
International Criminal Court	International criminal investigations
International Criminal Court	Victims' participation and reparations
International Criminal Court	International law, international relations, public administration, and change management
University of Bern	Organisational studies, anthropology of the state, legal anthropology, and sociology of quantification
Kosovo Specialist Chambers	International criminal law
Embassy of the Federal Republic of Germany	International criminal law
Supervisory Board of The Hague Institute for Innovation of Law	Economics, budget planning, financial management, strategic guidance, information management, and international criminal justice

Annex 3: Interview Participants

Institutional Affiliation	Area of Expertise
(formerly) Open Society Foundations and Harvard University	Criminal justice indicators, international criminal law
University of Amsterdam	International criminal law, international criminal procedure
PluriCourts, University of Oslo	Public international law, performance of international courts, ECtHR
Centre for International Law Research and Policy (CILRAP)	International criminal law
The John Marshall Law School	International criminal law, international criminal courts
Kosovo Specialist Chambers and Specialist Prosecutor's Office	Judicial administration, criminal law and procedure, international criminal justice, humanitarian law and human rights
Victoria Law Foundation	Empirical legal studies, social statistics
International Criminal Court	Financial planning and control, budget forecasting, risk management
Pontificia Universidad Javeriana and Oxford University	Comparative transitional justice mechanisms; empirical legal research methods
Dejusticia	Transitional justice, international criminal law, and empirical legal research

Annex 4: Estimated Timeline and Budget

It is anticipated that a project of this magnitude would evolve in stages. It may be conceived from the outset in two ways:

- A. **Comprehensive scope**, *i.e.*, with the ultimate intention to achieve full or near full coverage of all major ICJ situations within a certain time, or
- B. **Limited scope**, *i.e.*, with the ultimate aim of covering only a handful of these situations, for demonstrative purposes.

In both scenarios, the activities and data collection methods would be similar (explained in detail above):

- Expert questionnaires;
- Probability sample polling of the general public for a very limited number of questions;
- Non-probability sampling of the general public for a larger number of questions;
- Non-probability sampling of specific populations;
- Focus groups and in-depth qualitative interviews;
- Collection and analysis of existing data from governments, NGOs, international organisations, etc., about the situation;
- Collection and analysis of data from the ICC and other relevant mechanisms (e.g., the Colombian Peace Court).

Option A: Comprehensive Scope

The benchmarks' ultimate intended coverage would include:

- Situations currently under ICC-OTP preliminary examination activities.
- Situations under ICC investigation.
- Cases currently at trial stage and reparations stage at the ICC, excluding closed cases.
- Other situations where international crimes appear to have been committed in the past 5 years.

For these situations, the benchmarks would measure the effectiveness of the actions taken (or not taken) by key ICJ actors, with respect to various goals, at different stages of the process. While it would be desirable to measure all these situations in full (with all the elements of the methodology), it is anticipated that security concerns and other factors will restrict the ability to collect some of the necessary data in some of these countries. In these cases, a partial deployment of the methodology may be an acceptable alternative (for instance, including only expert questionnaires, qualitative interviews and data from the ICC and other international organisations, and excluding general population surveys and focus groups).

Methodological choices about data collection methods and levels of coverage will be in part driven and adjusted based on contextual considerations. Given that the subject matter of international criminal justice includes a wide diversity of situations, the methodology for the Nuremberg Benchmarks does not aim at full standardisation of data collection methods and coverage in all countries (such as in the World Justice Project's Rule of Law Index). Some data may be available for many situations, such as approximate numbers of people killed or forcibly displaced and number of offenders prosecuted. Other data can only be collected by interviewing survivors in the particular situation.

Option B: Limited Scope

The benchmarks' coverage would be restricted to a handful of situations that may act as demonstrative examples. Evolution in these situations would be tracked yearly, for a period of 5 years. This limited scope option would aim at developing evidence for multiple stakeholders of the potential impact of the benchmarks.

Subject to multiple assumptions and choices listed throughout this document, an illustrative timeline for this project would be as follows:

Stage	Option A. Comprehensive coverage (20 situations)		Option B. Limited scope (4 situations)	
	Duration (months)	Approx. Cost (euro)	Duration (months)	Approx. Cost (euro)
Phase 1. Development of the methodology	8	100,000	4	50,000
Phase 2. Pilot Benchmarks for the current situations in Colombia and Bangladesh/Myanmar, from multiple perspectives (including measurement of ICC and local actors).	14	350,000	10	242,000
Phase 3. Full deployment of the methodology	30	2,000,000	14	400,000
Phase 4. Yearly update of the Benchmarks	12	950,000	12	200,000

The same stages (e.g., development of the methodology; piloting in two situations), would be more time consuming and costly in Option A than in Option B, because of the need to adjust the work to the ultimate intended outcome of the whole project. In other words, piloting the methodology in one or two situations, for future deployment in 20 situations, is far more complex than piloting the same methodology for future deployment in a handful (3 or 4) situations.

Based on experience from previous indicators (World Bank, World Justice Project, HiiL, etc.), it is anticipated that in both Options A and B, a significant decrease in costs would occur for the yearly update of the benchmarks, beyond final deployment of the full methodology. These savings not only come from a learning curve, but also from methodological shortcuts that may be implemented beyond year 1, without compromising the overall academic rigour of the project; for instance, a gradual shift from relatively more expensive probability sampling in year 1, to less expensive non-probability sampling in years 2, 3 and 4 (similar to the national Census methodologies, which in most cases provide for full data collection every 10 or so years, and partial data collection in intermediate years).

Contact information

Maurits Barendrecht, Programme Director NL
The Hague Institute for Innovation of Law (HiIL)
T: +31 (0) 70 762 0700
E: maurits.barendrecht@hiil.org
Muzenstraat 120, 2511 WB The Hague, The Netherlands
www.hiil.org

Jolana Makraiová, Senior Officer for Interdisciplinary
Research
International Nuremberg Principles Academy
T: +49 911 14 89 77 - 26
E: Jolana.Makraiova@nurembergacademy.org
Bärenschanzstrasse 72, 90429 Nuremberg, Germany
www.nurembergacademy.org

