Constructing and validating an instrument for comparing national criminal justice policies

Construcción y validación de un instrumento para comparar las políticas nacionales de justicia penal

Construindo e validando um instrumento para comparar políticas nacionais de justiça criminal

Date of receipt: 2020/05/07 | Date of evaluation: 2021/03/24 | Date of Approval: 2021/04/15

Lorea Arenas-García
PhD in Criminology
Lecturer in Criminology, Department of Public Law
Universidad de Extremadura
Cáceres, España
lorea@unex.es
ORCID: https://orcid.org/0000-0002-4997-9163

Para citar este artículo / To reference this article / Para citar este artigo: Arenas-García, L. (2021). Constructing and validating an instrument for comparing national criminal justice policies. *Revista Criminalidad, 63*(3), 107-125. https://doi.org/10.47741/17943108.313

Abstract

The social exclusion generated by criminal justice policies adopted by some countries regarding certain individuals (offenders, ex-offenders, defendants, or suspects) places these individuals in worse individual and social conditions pursuant to their contact with penal institutions. Measuring social exclusion is useful for comparing different national crime control systems and for estimating how exclusionary a country is. However, traditional comparisons of criminal justice policies have focused on the level of punitiveness exerted by each national criminal justice system referring to the incarceration rate per 100.000 inhabitants. As pointed out by Díez-Ripollés y García-España (2019), constituting all comparisons around one indicator marginalises others with stronger measurement capacity. For this reason, an instrument of 39 indicators was created by the authors using expert judgments (Díez-Ripollés & García-España, 2019). This article discusses the development of the tool in detail and how its validity and reliability were established. To validate the instrument, we surveyed 99 international experts and applied Aiken V as the main statistical procedure. Our findings suggest this instrument is valid and reliable for international comparison.

Keywords

Criminal justice, law enforcement, (source: Criminological Thesaurus - United Nations Interregional Crime and Justice Research Institute (UNICRI). Comparative analysis, research methods, indicators, questionnaires. (source: author).

Resumen

La exclusión social generada por las políticas de justicia penal adoptadas por algunos países a ciertas personas (delincuentes, exdelincuentes, acusados o sospechosos), los deja en peores condiciones individuales y sociales después de su contacto con las instituciones penitenciarias. La medición de la exclusión social es útil para comparar los diferentes sistemas nacionales de control de la delincuencia y estimar qué tan excluyente es un país. Sin embargo, las comparaciones tradicionales de la política de justicia penal se han centrado en el grado de punitividad ejercido por cada sistema nacional de justicia penal en referencia a la tasa de encarcelamiento por cada 100000 habitantes. Como señalan Díez-Ripollés y García-España (2019), construir todas las comparaciones en torno a un indicador margina a otros con mayor capacidad de medición. Por este motivo, los autores crearon un instrumento de 39 indicadores mediante el juicio de expertos (Díez-Ripollés & García-España, 2019). Este artículo discute en detalle cómo se desarrolló la herramienta y cómo se estableció su validez y confiabilidad. Para validar el instrumento, encuestamos a 99 expertos internacionales y aplicamos Aiken V como el principal procedimiento estadístico. Nuestros hallazgos sugieren que este instrumento es válido y confiable para la comparación internacional.

Palabras clave

Justicia penal, aplicación de la ley (fuente: Tesauro Criminológico-Instituto de Investigación Interregional de Crimen y Justicia de las Naciones Unidas-Unicri). Métodos de investigación, indicadores, cuestionarios, análisis comparativo (fuente: autor).

Resumo

A exclusão social gerada pelas políticas de justiça criminal adotadas por alguns países a certas pessoas (criminosos, ex-infratores, acusados ou suspeitos) os coloca em piores condições individuais e sociais após seu contato com instituições penitenciárias. Medir a exclusão social é útil para comparar diferentes sistemas nacionais de controle de crimes e estimar o quão excludente é um país. No entanto, as comparações tradicionais da política de justiça criminal têm se concentrado no nível de punição exercido por cada sistema nacional de justiça criminal em referência à taxa de encarceramento por 100.000 habitantes. Como apontam Díez-Ripollés y García-España (2019), construir todas as comparações em torno de um indicador marginaliza outros com maior capacidade de medição. Por essa razão, os autores criaram um instrumento de 39 indicadores através do julgamento de especialistas (Díez-Ripollés & García-España, 2019). Este artigo discute detalhadamente como a ferramenta foi desenvolvida e como sua validade e confiabilidade foram estabelecidas. Para validar o instrumento, pesquisamos 99 especialistas internacionais e aplicamos a Aiken V como principal procedimento estatístico. Nossos achados sugerem que este instrumento é válido e confiável para comparação internacional.

Palavras-chave

Justiça criminal, aplicação da lei. (fonte: Thesaurus Criminológico - Instituto inter-regional de pesquisa das Nações Unidas para o Crime e Justiça - UNICRI). Métodos de pesquisa, indicadores, questionários, análise comparativa. (fonte: autor).

Introduction

The comparison of different crime control systems is a complex task. Comparative criminal justice policy has traditionally focused its analysis on the level of punitiveness exerted by each national crime system, presupposing certain humanitarian limits in the enforcement of penalties and using exceedingly limited indicators of punitive moderation. As pointed out by Díez-Ripollés & García-España (2019), the traditional approach is impaired since nearly all comparisons are basically based on the incarceration rate per 100,000 inhabitants. Such circumstance marginalises other indicators with a strong capacity for expression, such as, for example, the number of criminal proceedings that end in conviction or the average length of imposed sentences.

As a consequence, a comparison of international criminal justice policy should begin with a more enriched and comprehensive approach, using a broad set of indicators for this purpose (Díez-Ripollés & García-España, 2019). The methodology that operationalises generic or abstract concepts in empirical and observable indicators in order to establish measurement instruments is used in many fields of knowledge, with different purposes. For example, in Sociology scales that verify public opinion on certain phenomena are quite common; in Psychology, the creation of tests for evaluating personality and behavioural features is widely used; and those instruments are also used in natural sciences, as many indicators have been formulated in terms of them. Said measurement instruments are, in turn, applied to different sociodemographic contexts in order to carry out international comparisons. Proof of that would be the Social Progress Index (SPI) or the World Health Organisation (WHO) database.

In the criminal justice policy arena, there are several secondary data sources that enable comparison among countries by means of using basic indicators, such as the incarceration rate, the overcrowding rate, the number of alternative measures to imprisonment, etc. Notable among those data sources is the *World Prison Brief* (on the global level) and the "SPACE" annual statistical reports from the Council of Europe (on the European level). Such information has been traditionally used for identifying criminal justice policy trends among countries in rigoristic terms, even though there is no instrument to empirically verify those trends; thus, the importance of establishing a set of valid and reliable indicators to

enable generalisation. In order to address this need, Díez-Ripollés and his team have developed several research projects[1] that are groundbreaking in the creation, and subsequent validation of an instrument for international comparison that allows to classify national crime control systems according to their socially exclusionary effects.

This work describes the methodology adopted in such projects and addresses the following aspects: the baseline theoretical framework, the justification of the method adopted for each research objective (creating and validating a measurement instrument), the measurement instrument implementation phases, the relevance of the statistical tests used, as well as the limitations that were contended.

Theoretical framework

The research developed by Díez-Ripollés is based on a specific theoretical framework (Díez-Ripollés, 2011; 2013) which highlights the effects of the crime control system on individuals and groups who are priority targets of the criminal prevention and prosecuting bodies, namely offenders, ex-offenders, defendants or suspects. For the author, the quest for either social exclusion or social inclusion undertaken by those individuals prone to enter into conflict with the criminal law reflects two contrasting approaches to the criminal justice policy objective of preventing crime. The *socially exclusionary* approach is essentially aimed at achieving the incapacitation of the groups referred to, which implies ensuring that offenders, ex-offenders, defendants or suspects, —after their contact with penal institutions—, are set in individual and social conditions where it will be more difficult for them to break the law or to avoid being discovered. Conversely, a *socially inclusive* approach seeks, above all, the social reintegration of such groups, so that offenders, ex-offenders, defendants or suspects, —after their contact with penal institutions—, establish themselves in the same or better individual and social conditions in order to voluntarily lead a law-abiding life.

Therefore, the different national crime control systems[2] will have to be evaluated according to the greater or lesser adherence of their penal intervention models to one of the two approaches, even though the author has focused his research, more modestly, on a strict socially exclusionary approach and has identified nine major areas of penal intervention which are especially adequate to show the relevant exclusionary effects that, as a whole, offer a comprehensive image of the corresponding criminal justice system. The nine dimensions, also called "pools", are as follows[3]: control of public spaces (gated communities, video-surveillance, restriction of access to public spaces); legal safeguards (undermining of due process safeguards, hindering or restriction of legal remedies); sentencing and sanction systems (judicial discretion, aggravated provisions for recidivists, extensive use of prison, alternative sanctions to imprisonment, electronic monitoring); harshest penalties (death penalty, life imprisonment, long-term prison sentences); prison rules (living conditions in prison, respect for prisoners' rights, release on parole); preventive intervention (pre-trial detention, indefinite preventive detention); legal and social status of offenders and ex-offenders (disenfranchisement, deprivation of additional civil rights, accessibility to social resources); police and criminal records (extension and accessibility of records, community notifications); and youth criminal justice (age thresholds, treatment differentiated from adults).

Creation of the instrument

The socially exclusionary theoretical approach is a theoretical definition that lacks the necessary accuracy to measure the phenomena which it refers to. It is a construct or abstract idea that cannot be directly observed, the function of which is essentially synthetic. In other words, to determine if a country is closer to or farther away from the socially exclusionary approach, it is necessary to operationalise its content, translating it into empiric variables or measurable indicators.In this context, implementing the method for the creation of indicators is justified by three essential reasons: the aim is to verify a particular theoretical model; there are no similar measurement instruments and this prevents the possibility of building the instrument based on others; and it does not preclude the integration of indicators traditionally used in the sphere of comparative criminal justice policy. The methodology for creating indicators requires them to be valid and reliable (Casas, 1989; González-Blasco, 1994; Lazarsfeld, 1985; Messick, 1989; Prieto & Delgado, 2010; Rey del Castillo, 2004; Sierra-Bravo, 2008).

1 *"The Evolution of Criminal Justice Policy in a World of Increasing Social Exclusion"* (DER2012-32070) and *"Measuring the Social Exclusion Generated by the Criminal Justice Policy in Western Developed Countries"* (DER2015-64846- P), both financed by the Ministry of Economy and Competitiveness.

2 The attention has not been focused on the study of the political-structural, socio-economic or cultural factors that promote the adoption of certain rules or practices in the different criminal control systems that give rise to more or less socially inclusive results in the sense pointed out in this work. For a more detailed explanation consult: Díez-Ripollés, 2011.

3 To discern why other theoretical dimensions have been excluded, see Díez-Ripollés, 2011.

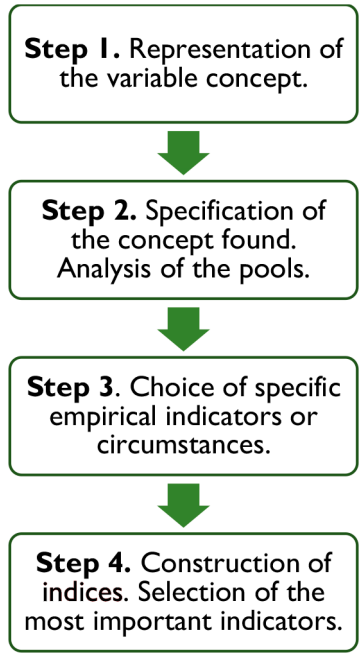
Messick (1989), whose theory on the validity of tests has had a deep bearing on the literature, defined "validity" as follows: "Validity is a global evaluative judgment of the degree to which empirical data and theoretical logic support the conception and convenience of the inferences and actions carried out based on scores produced by tests or other measuring instruments" (p. 19).

Other authors describe it as "the gap between the concept and those indicators chosen for its representation" (Estévez-García & Pérez-García, 2007, p. 55) or "the justification that the items to measure criterion are a representative sample of the content to be evaluated" (Prieto & Delgado, 2010, p. 70). In turn, to measure validity it is essential to use a tripartite structure of *criterion, content and construct* (Hernández-Sampieri *et al.,* 2006).

Criterion validity is reached with the comparison of data obtained with others that have been previously used to measure the same concept. And this is only possible by taking into account indicators already validated in previous research, as would be the already mentioned incarceration rate per 100,000 inhabitants. Content validity has been traditionally defined by Nunnally (1973) and Mehrens & Lehmann (1982) as "the degree to which the items making up the sample represent the content subject that is being measured" (Escurra, 1988, p. 105). In other words, it examines whether the measurement includes the variety of meanings which such concept may adopt, whether they are relevant for the main theoretical construct, and it is therefore important to draw up a list with the largest possible number of indicators so they can subsequently be discarded. Regarding the validity of the construct, it refers to the theoretical or abstract ideas which are at the basis of the research and from which operational definitions arise. If the latter properly reflect the theoretical meaning of a concept, the construct's validity will be greater. In our case, the construct is the theoretical definition of social exclusion arising from a larger theoretical framework.

Likewise, the creation of indicators methodology requires them to be reliable, that is, to be able to obtain consistent results in successive measurements of the same phenomenon. In this sense, reliability refers to "the consistency of the measurement, the possibility of replicating the results obtained with an instrument if it is applied in different occasions" (Estévez-García & Pérez-García, 2007, p. 55).

As regards the procedure of variable operationalisation, it was developed by Lazarsfed (1985) who differentiates four successive steps (see graph below).



Graph 1. Procedure for variable operationalisation.

This procedure was adapted and implemented by the research team throughout the entire study. The first two steps were developed based on the concept that had been theoretically substantiated by Díez-Ripollés in previous research. The author had previously set forth a theoretical notion defining the main features of exclusion and its pools. On the basis of this theoretical circuit and in order to comply with the variable operationalising procedure, a group of researchers specialised in penal law, criminology and criminal justice policy was set up (phase 1) so that they may assess the theoretical concept of the project (phase 2: steps 1 and 2), by carrying out its operationalisation by means of the creation of indicators classified in different dimensions of the exclusion (phase 3: step 3), in order to end with a rigorous selection of them (phase 4: step 4).

Setting up a group of researchers (Phase 1)

The fact of having an expert judgment offers great advantages when establishing knowledge. An expert is:

> An individual, group of individuals or bodies that can offer maximum competence, conclusive assessments regarding a given problem; make real and objective forecasts on the effect, applicability, feasibility and relevance which the proposed

> solution may have in practice; and provide recommendations on what may be done to improve it. (Crespo, 2007, p. 7)

Experts provide a greater level of profundity in the matter; their responses are detailed and their contents have a high theoretical quality (Cabero-Almenara & Barroso-Osuna, 2013). That is why their relevance for creating indicators is justified. In this first phase, there were 20 researchers assigned to the research project and it was assumed that they were experts in the matter. In order to determine this aspect with objective criteria, the Delphi method was partially applied (Cruz, 2009). This method has been quite frequently used in the last few years for individual self-assessment, in order to recognise their expertise competence (Blasco *et al.,* 2010; García & Fernández, 2008; López, 2008; Oñate, 2001).

Said method gives validity to the methodological process by confirming the expert level of the research group as regards each dimension or pool, by allowing to establish criteria for the inclusion or exclusion of experts. It specifically assesses five different argument sources, namely: knowledge of the problem's current status, understanding the problem, capacity to analyse the problem, experience in theoretical research development and experience in empirical research development. In order to implement the Delphi protocol, a brief three-page questionnaire was prepared and sent on line to each team member, so as to know their individual and professional characteristics, as well as their competence or expertise level in each theoretical dimension or pool, specifically, the expert had to score on a 1-to-5 Likert scale whether their level of expertise was very low (1), low (2), medium (3), high (4) or very high (5).

Concerning the statistical testing used, some statistical descriptions of the group were carried out in the first place (frequency, averages, minimum and maximum range) in order to subsequently develop the mathematical procedure described in the Delphi methodology, which allows for acceptance or rejection of the researcher, in accord with his or her own evaluation, thanks to the calculation of the K coefficient or the expert knowledge coefficient.

The first descriptive results showed that, for the most part, researchers attained medium and high scores. As seen on the table below, they were classified in four categories and ranges according to the average score obtained in the pools: very low (1-1.50), low (1.51-2.50), medium (2.51-3.50), high (3.51-4.50) and very high (>4.50). Only researcher no. 14 was classified within the "very high" score category by achieving the highest average, over 4.50 (4.71 points), while also being the expert with the greatest number of publications. The second group includes 5 researchers in the "high" category, composed of those numbered 2, 10, 12, 20 and 5. Their averages ranged between 3.51 and 4.50. We then have a third group, which is larger and includes 12 researchers in the "medium" category, with an average between 2.51 and 3.50 points. And a final group of two researchers (numbered 1 and 19) with a "low" level and averages between 1.51 and 2.50 points. It was similarly determined that pool number 5 had the highest score, with 355 points: the highest score which a pool could obtain was 500 points (if all judges granted 5 points). It was also the pool with the highest score with reference to the following argument sources: "knowledge of the problem's current status", "understanding the problem" and "capacity to analyse the problem". This was followed by pool 3 (340 points), 4 (337 points) and 2 (334). Finally, it was found that researchers had more experience in developing theoretical research in pools 2 and 3, and more empirical experience in pools 3 and 5.

Table 1.

Group statistical descriptions according to maximum scoring in the pool, average, range and category.

Researcher (N°)	High score	Total pool average	Range	Category
14	212	4.71	> 4.50	Very high
2	189	4.20	3.51 – 4.50	High
10	173	3.84	3.51 – 4.50	High
12	173	3.84	3.51 – 4.50	High
20	162	3.60	3.51 – 4.50	High
5	160	3.56	3.51 – 4.50	High
4	148	3.29	2.51 – 3.50	Medium
13	144	3.20	2.51 – 3.50	Medium
17	142	3.16	2.51 – 3.50	Medium
11	138	3.07	2.51 – 3.50	Medium
3	136	3.02	2.51 – 3.50	Medium
8	133	2.96	2.51 – 3.50	Medium
9	132	2.93	2.51 – 3.50	Medium
6	127	2.82	2.51 – 3.50	Medium
7	125	2.78	2.51 – 3.50	Medium
18	124	2.76	2.51 – 3.50	Medium
15	119	2.64	2.51 – 3.50	Medium
16	117	2.60	2.51 – 3.50	Medium
1	107	2.38	1.51 – 2.50	Low
19	105	2.33	1.51 – 2.50	Low

With respect to the determination of the K coefficient or the expert level, it was carried out by calculating the subtotal addition of the knowledge level coefficient on the subject under study (kc) and the average of argument sources (Ka), in other words, K=1/2 (Kc + Ka). The coefficient has a minimum value of 0 and a maximum of 1, considered as a high level of competence when the coefficient is equal to or higher than 0.75. It is noted that on Table 2 there are a minimum of 4 researchers with a high level of competence (K=0.75 or >) in each pool and a maximum of 11. Categorising the number of researchers with a high level of competence from lowest to highest, it is noted that: the first pool includes 4 researchers; the sixth pool includes 5; the eighth, 6; in the second and third pools there are 7; in the ninth pool there are 8 researchers; in the fourth, 10 and in the fifth, 11.

Finally, coefficient K was calculated at group level, and it was determined that it stood at 0.55 points, an average score which revealed that the team did not consider itself an expert in all the issues covered by the questionnaire and that, even if there were some individuals deemed experts in more fields and others in fewer fields, the team as a whole did not achieve a high level of competence, possibly due to the presence of researchers in early training stages. Notwithstanding, this aspect was not necessarily taken as negative or restricting, since it allowed for the adoption of future methodological decisions, for instance the fact of ensuring the presence of researchers —one in each pool— with a high level of competence in the nine working groups created *ad hoc* in the subsequent creation of indicators.

Theoretical assessment of the concept (Phase 2)

Another essential aspect of the instrument creation process was verifying whether all members of the research team had properly understood the theoretical concept upon which all the indicators would be established or operationalised. Although it was a concept that had already been theoretically substantiated by the author (Díez-Ripollés, 2011; 2013), the team was asked whether it was precise and clear, by means of an e-mail message that was

Table 2.
Coefficient K per researcher and pool.

K	P1	P2	P3	P4	P5	P6	P7	P8	P9
R1	0.77	0.3	0.3	0.53	0.53	0.52	0.49	0.53	0.52
R2	0.71	0.54	0.96	0.96	0.96	0.96	0.96	0.88	0.85
R3	0.4	0.91	0.94	0.87	0.69	0.52	0.74	0.4	0.52
R4	0.72	0.91	0.72	0.87	0.94	0.69	0.56	0.4	0.72
R5	0.85	1	0.83	0.78	0.83	0.83	0.83	0.92	0.45
R6	0.67	0.43	0.64	0.51	0.74	0.39	0.72	0.24	0.78
R7	0.52	0.76	0.74	0.74	0.76	0.52	0.52	0.52	0.52
R8	0.71	0.74	0.56	0.74	0.76	0.37	0.76	0.56	0.54
R9	0.69	0.91	0.71	0.87	0.56	0.37	0.56	0.37	0.73
R10	0.64	0.64	0.64	0.64	0.82	0.64	0.76	0.98	1
R11	0.71	0.73	0.72	0.73	0.72	0.54	0.7	0.6	0.67
R12	0.81	0.71	0.96	0.83	0.96	0.76	0.76	0.84	0.8
R13	0.72	0.72	0.72	0.87	0.72	0.94	0.56	0.4	0.72
R14	1	0.98	0.98	0.98	0.98	0.96	1	0.9	0.96
R15	0.67	0.37	0.6	0.53	0.76	0.35	0.53	0.72	0.7
R16	0.52	0.76	0.49	0.74	0.59	0.38	0.69	0.56	0.4
R17	0.56	0.73	0.8	0.69	0.58	0.56	0.56	0.54	1
R18	0.49	0.69	0.76	0.35	0.78	0.47	0.35	0.78	0.76
R19	0.35	0.52	0.71	0.92	0.52	0.35	0.54	0.35	0.49
R20	0.58	0.74	0.58	0.76	0.98	0.73	0.89	0.58	0.98

individually and collectively answered by the research team members. 65% of the researchers (13 out of 20) responded by asking something else, 6 individually and 7 collectively; the remaining 35% did not raise any objection to the proposed concept. In spite of being considered a concise and clear definition, it led to doubts regarding: whether it was a too optimistic definition, since it aimed to prevent the subject from committing a crime or to avoid being discovered; if this were the case, if such goals were achieved, anybody would accept a socially exclusionary control system; vague with regards to the "materially" concept; exceedingly normative; and confusing as to whether the exclusionary character derives from adopting measures that make reintegration of the subject into society difficult or impossible, or else from the mere imposition of a criminal penalty. It was concluded that the exclusionary character was caused by adopting measures that hindered the reintegration of the subject into society, thus forcing those in conflict with penal law to follow the path of crime, since they do not have the possibility of being reintegrated into society. This is why the criminal sanction is identified with the simple exclusion from society.

Successively, all doubts —after being systematised— were discussed and clarified in a group meeting, until consensus was reached.

Operationalisation of the concept broken down into pools (Phase 3)

Once the theoretical doubts had been cleared up, the concept was operationalised so that it could be measured in the nine exclusion pools that have already been mentioned. Each one of them would include a significant number of indicators in order to cover all aspects of the dimension, either as real punitive rules or practices that are effectively applied or whose implementation is considered in the Western developed countries. An indicator formulated as a rule represents a legal standard which usually, though not always, is included in penal law and establishes certain consequences for certain behaviours or situations related to crime control, for instance the enshrinement of the death penalty in the Criminal Code. Contrastingly, punitive practices measure the way in which the different social agencies effectively react to behaviours or situations related to crime control, whether or not in agreement with the provisions established by law. For example: applying the death penalty to individuals belonging to certain ethnic minorities to a much greater extent than to other groups.

In order to facilitate the creation of indicators, the coordinating group divided the research group into nine working subgroups composed of 2 or 3 individuals, according to the specialisation and expertise attested at each dimension, and established the methodological criteria to build those indicators so that they might truly represent the exclusion concept; in other words, be valid. And these were: (1) the diverse rules or practices chosen for each pool should be drafted to admit only dichotomous replies ("Yes" or "No"); (2) an effort should be made to include a good number of rules and practices — so that subsequently the best could be chosen—, they should be as expressive as possible without there being any overlapping among them; (3) the indicators should be drafted in the same sense, that is, with the "yes" reply signifying exclusion, and the "no", inclusion; (4) an effort should be made to include both rules and practices in each pool; (5) and they had to be drafted in English.

Likewise, the coordinating team established a schedule of successive working meetings for groups and subgroups. The former, a total of 18, underwent a critical review of the rules and practices proposed for each pool by each subgroup, and the go-ahead to those finally selected was given through the assent of the whole team. General meetings were established as a time for reflection, in order to reinterpret —from different perspectives— each one of the items that were being put together.

The work resulted in the identification of 278 punitive rules and practices distributed among the nine groups.

Selection of indicators (phase 4)

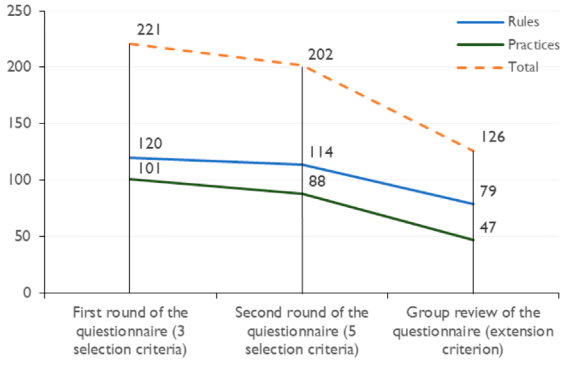
Taking into account the large volume of resulting indicators, it was agreed that they would be reduced in order to obtain a handy and manageable instrument. It was decided that the way to carry this out would be through the implementation of 3 indicator selection criteria, assessing: *completeness*, which expresses the special relevance of the information supplied or its importance; *extension*, which ensures that the information does not deal with aspects that are exceedingly particular or specific; and *ease of access*, whereby the information is expected to be obtained relatively simply.

This task was only carried out by the four members of the coordinating team individually. They responded to a questionnaire that was designed for such purpose, scoring 1 when the criterion was fulfiled, and 0 when it could not be deemed fulfiled.

Each indicator could obtain a maximum score of 12 points, if all team members gave 1 point to each one of the three criteria specified.

Once the first questionnaire had been sent it was decided to discard those indicators with 4 points or less in view of the results, thus reducing the instrument to 221 indicators. Then another session was carried out, which improved the selection criteria by introducing two new ones: the capacity to discriminate among countries and the clarity of the wording, expressly so that the indicator be easily understood.

Once the results from this round were obtained, those items with scores of 4 and 3 points were selected, and the items with 1 or 0 points in completeness or in discrimination were removed, since the team considered that these two criteria were essential. Thanks to this procedure, the instrument was reduced to 202 items. Thereafter and for a period of one month, the coordinating team reviewed all the pools in successive group meetings, starting with those with a larger number of indicators. The goal was to undertake a detailed review of the extension of those items with accumulated scoring that had not been excluded in the previously described situations, but that could be removed or recast since they dealt with topics that were repeated or too specific. Said task was carried out while at the same time seeking a balance in the distribution of rules and practices. As shown on the graph below, once the selection process was completed, 76 items were discarded, many of them from pools 1, 5, 6 and 7, thus obtaining an instrument consisting of 126 items made up of 79 rules (63%) and 47 practices (37%).



Graph 2. Number of rules and practices according to selection procedures and criteria

It should be noted that throughout the whole 'cleaning' process, the number of rules was always higher than the number of practices, even though the progressive reduction of both maintained a certain proportion. As far as the final distribution of rules and practices in each pool (see Table 3), as was to be expected, for most of the dimensions, the rules were higher —in percentage terms— than the practices, particularly in pools 2 and 7, where the gap was greater, as opposed to pools 6 and 8, where the distribution was 50/50.

Table 3.
Number of rules and practices selected per pool.

Pools	Rules	%	Practices	%	Total	%
Pool 1	8	66.67	4	33.33	12	100
Pool 2	11	73.33	4	26.67	15	100
Pool 3	9	64.29	5	35.71	14	100
Pool 4	7	53.85	6	46.15	13	100
Pool 5	13	68.42	6	31.58	19	100
Pool 6	6	50.00	6	50.00	12	100
Pool 7	9	75.00	3	25.00	12	100
Pool 8	7	50.00	7	50.00	14	100
Pool 9	9	60.00	6	40.00	15	100
Total	**79**	**63**	47	37	126	100

Validation of the instrument

Once the instrument was established with 126 indicators, the content validation process was initiated through the judgment of international experts. Prior to this, the instrument was translated into English and then reviewed by a Spanish-speaking British woman who specialised in law and criminology.

Content validity is a methodology which is widely used to ascertain whether the indicators of a measurement instrument are relevant and representative of the baseline theoretical construct (Ding & Hershberger, 2002). Such validation is usually carried out through the judgment of experts, in other words, by way of the assessment carried out by very qualified and reputed scientific and professional individuals. In our case, it was decided to focus on the

individual judgments of experts, where each expert validated the instrument's items individually[4].

For this task to be carried out, a rigorous procedure was undertaken and broken down into four stages. This allowed to arrange the information so as to operationalise and achieve efficiency in the process, specifically: selection and contacting of international experts (stage 1), design and online management of the instrument (stage 2), statistical analysis of results (stage 3) and final selection of the strongest indicators (stage 4).

Selection and contacting of international experts (Stage 1)

Selecting the experts that will take part in the validation of the instrument is an essential and critical aspect of the methodological process (Escobar-Pérez & Cuervo-Martínez, 2008, pp. 28-29), since "the quality of the results obtained through a study where the judgment of experts is applied will be entirely related to the experts chosen" (Cabero-Almenara & Barroso-Osuna, 2013, p. 28). Skjong y Wentworth (2000) list a set of selection criteria that should be taken into account: academic and practical background, reputation within the community, motivation and availability to participate, as well as impartiality. At the same time, the methodological requirements of targeted qualitative sampling added another two criteria: achievement of a theoretical saturation thanks to the creation of prototype expert categories; and geographical representativeness, since the instrument was designed to be implemented in the Western developed countries.

Taking such methodological requirements into account, the objective was to choose experts from 20 countries in Central Europe, Eastern Europe, Northern Europe, Southern Europe and English-speaking countries in and outside of Europe. Finally, the participation of 18 of them was achieved (except for Austria and Sweden), namely: Belgium, Canada, Denmark, Finland, France, Germany, Hungary, Ireland, Italy, the Netherlands, New Zealand, Poland, Portugal, Romania, Spain, Switzerland, United Kingdom and the United States.

Efforts were also made to ensure that experts were academics and professionals with a doctorate (PhD)[5] and extensive knowledge and experience (over 10 years) in the field of crime prevention or prosecution and to have a balance —as far as possible— between academics and professionals, gender and political orientation; and to have diversity according to the pool's specialisation.

Methodological debate on the adequate number of experts for validation by way of the judgment of experts is very varied. Some authors suggest a range from 2 to 20 experts (Gable & Wolf, 1993; Grant & Davis, 1997), while others suggest a lower amount of approximately 10 (Hyrkäs *et al.,* 2003). In our case we decided to increase considerably the number of experts, so as to facilitate management of a very wide-ranging instrument, achieve the saturation of the experts' prototype and obtain an adequate geographical representation.

The targeted sampling was carried out smoothly since we engaged individuals with a great reputation in the academic and/or professional fields, whose experience by and large exceeded 20 years. The coordinating team contacted those experts who met the requirements and requested their participation in the project. This was carried out by means of an e-mail with an attached circular letter including the necessary information with regard to their performance within the research. If they agreed to participate, the aforementioned letter described two key required actions: to supply the name of three experts that were motivated and available for the validation of the questionnaire; and to commit to validate the questionnaire, once the group of experts formed by the provided names had done so. In this way, the first group of experts facilitated the locating and selection of the second group of experts, by using a non-probabilistic and targeted snowball sampling, forasmuch as it was necessary to identify potential experts who, in the case of non-academic professionals, were more difficult to find. In other words, the sampling worked as a chain determined by the assistance that the first group of experts provided to the coordinating team in order to identify and locate the required profiles.

As the experts of the first group agreed to participate, a brief online questionnaire was sent to them in order to collect various personal and professional information to verify they effectively met the selection criteria. Thanks to this method 58 experts were located, of which 28 agreed to participate. As

4 Note that, the objective is not to verify empirically that such indicators effectively generate social exclusion. Such verification is carried out through a broad consensus of experts.

5 Initially, the requisite of having a PhD was established for both groups (academic and professional individuals). However, during the selection process it was decided that the professional group would not be required to comply with said condition, in order to increase the possibilities of finding people with a highly practical profile.

regards the individual and professional features of the group, it was found that: 71% were men, a percentage which is consistent with the usual distribution of the senior academic population; they were, on average, 58 years old and had a standard deviation of 7.4 years, considering that at the time of responding to the questionnaire the youngest was 42 and the oldest, 71; they were predominantly of Spanish, German, Swiss, Polish, British, US, Italian, Portuguese, Belgian and Dutch origin (all with a percentage of 7.1); having a progressive political orientation (92.3%); teachers (85.7%) who had developed —or were developing— work in the academic field (71%); with the expert who had the broadest professional experience in the field encompassing 21.4 years (vs. 19.2 for the one in the professional field); whose research was of empirical and theoretical nature, and hencewas a predominantly mixed profile (75%); they were experts in more than one pool (90% of the cases), and specifically, 60.7% were expert in three or less, while 39.2%, in six or less and 3.5% in more than six; and the greatest specialisation focused on pools 9 and 4 (13.9% and 11.9%, respectively).

The first group supplied a number of experts somewhat lower than the 84 that were expected. After the refusal of several of them to collaborate, we had 59 experts conforming the second group. In order to comply with the pre-established quota, the coordinating team sought —usually by way of the internet—, experts with the required profile. By virtue of this technique, 12 new individuals were added to the previous 59, thus obtaining a participation of 84.5%, or 71 experts. As regards the individual and professional features of the second group, it was found that: the distribution of experts per gender was quite balanced (51% women and 49% men); the youngest was 33 years old and the oldest 65, with an average age of 44 years and a standard deviation of 8.4 years; the questionnaire was completed by experts from 18 countries; 79% of the experts were considered to be progressive; being teachers (71.8%); that had developed —or were developin— work in the academic field (50.7%), followed by those with mixed (40.8%) and professional (8.5%) profiles; with more work experience in the academic than in the professional field (an average of 14 years vs. 10 years); whose research was predominantly mixed or theoretical-practical (63.4%). They were specialists in more than one pool: particularly, 81.4% were specialised in three pools or less; and 18.5% of them in six pools or less, especially in pools 3 and 6 (22.9% and 12.8%, respectively).

Table 4.

Selection and inclusion of experts in prototypical categories.

CONSERVATIVE PROFILE

1 Academic Experience in theorical reseach Expert (in one or several pools) Male/female	2 Academic Experience in theorical and empirical research Expert (in one or several pools) Male/female	3 Academic Experience in empirical research Expert (in one or several pools) Male/female
4 Mixed Experience in theorical research Expert (in one or several pools) Male/female	5 Mixed Experience in theorical and empirical research Expert (in one or several pools) Male/female	6 Mixed Experience in a empirical research Expert (in one or several pools) Male/female
7 Professional Experience in theorical reseach Expert (in one or several pools) Male/female	8 Professional Experience in theorical and empirical research Expert (in one or several pools) Male/female	9 Professional Experience in empirical research Expert (in one or several pools) Male/female

PROGRESSIVE PROFILE

10 Academic Experience in theorical reseach Expert (in one or several pools) Male/female	11 Academic Experience in theorical and empirical research Expert (in one or several pools) Male/female	12 Academic Experience in empirical research Expert (in one or several pools) Male/female
13 Mixed Experience in theorical research Expert (in one or several pools) Male/female	14 Mixed Experience in theorical and empirical research Expert (in one or several pools) Male/female	15 Mixed Experience in empirical research Expert (in one or several pools) Male/female
16 Professional Experience in theorical reseach Expert (in one or several pools) Male/female	17 Professional Experience in theorical and empirical research Expert (in one or several pools) Male/female	18 Professional Experience in empirical research Expert (in one or several pools) Male/female

Once the personal and professional characteristics of the selected groups had been analysed, the information was incorporated into the prototypical categories predefined by the team. This is what the literature calls the expert's 'biogram' profile (Cabero-Almenara & Barroso-Osuna, 2013, p. 28) and its prior establishment reduces the possibility of introducing methodological biases in the validation of the instrument, given that an indicator may be validated, at the same time, by two experts with completely different profiles, thus achieving greater reliability and saturation of the sample.

With this is mind, 18 categories were set up, as shown on Table 4, and divided into two axles of political orientation: conservative or progressive. In each of them, several things were considered: the expert's academic, professional or mixed character; their experience in theoretical or empirical research, or both; their specialisation as regards the pool and their gender. The criteria for including an expert in the categories established *ad hoc* were implemented by taking into account whether or not the case met the condition. The field of specialisation and the gender were not included in the formula and were complementary in the count. Nevertheless, they were subsequently taken into account to balance the sample, allowing for an adequate distribution thereof.

In the first group, category 11 —corresponding to a progressive academic profile, with experience in theoretical and empirical research— was the most represented, with 50% of the cases, followed by category 14 (25%) and category 12 (14.2%). Categories 1, 3 and 18 included only one expert each (3.6% respectively), and in the remainder there was no representation. Again, in the second group, category 11 was the most prevalent (28.2%), followed by category 14 (21.1%), 12 (8.5%) and categories 15 and 5 (7%). Presence was somewhat reduced in categories 3 and 13 (5.6% and 4.2%, respectively), followed by categories 10 and 17 (2.8%) and, very occasionally (1.4%), in categories 1, 2, 6, 8 and 18. Some experts (5.6%) could not be classified in any category, since their political orientation or the type of research they performed were unascertained.

Design and online management of the instrument (Stage 2)

Content validation was carried out in compliance with a series of methodological criteria, both in the design or planning of its content and in its management. The design of the instrument, the determination of the number of questions, their order and drafting were implemented taking into account the following guidelines (Sierra-Bravo, 2008): the inclusion of a pre-established number of questions to prevent it from being exceedingly long (i); that such questions be arranged in a logical and visually attractive manner (ii); involving very specific procedural instructions (iii); and the items of which would be set out at random (iv).

(i) Including a reduced number of items

An exceedingly long instrument can be tiresome and monotonous when filling it out (Sierra-Bravo, 2008). This reduces the motivation of the respondent and has a negative influence on the response rate. In general terms, no more than half an hour should be needed to complete it (Fowler, 1993). In order to resolve this situation, and taking the size of the baseline instrument (126 items) into account, the decision was made to implement it in two differentiated phases. In the first phase the instrument of 126 items —broken down into three parts consisting of 42 items—, was delivered exclusively to the second group, in particular, three questionnaire models were designed to be delivered to the three subgroups of experts resulting from the division of the second group. Thus, the second group, which was the largest, validated a higher number of items, which was more operational, proportioned and statistically consistent. In the second validation phase, the number of indicators to be validated —in this case by the first group of experts—, was reduced to 65, which was still a large number of indicators to be validated by a single expert. For this reason, two questionnaires were designed: one with 33 items and another with 32, and they were delivered to two subgroups of the first group, each containing 14 experts.

(ii) Arrangement of questions in a logical and visually attractive way

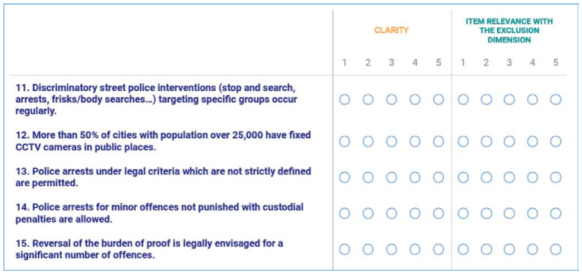
The questionnaires were structured in three parts. The first had two initial pages containing the objective of the research, its theoretical concept and pools, as well as the instructions for completion. The second part included the total number of indicators to be validated. Each item had to be answered before proceeding to the next page. Capital letters and colours were used to highlight certain key words, while the headings of each set of items were repeated every so often as a reminder of the scoring criteria. The third part was a brief questionnaire with personal and professional

questions (applied only to the second group of experts), followed by an open field for comments and a final message of appreciation (last page). Personal information was required following validation of the items, for two reasons: some questions could be embarrassing or difficult to answer (for example, those regarding age and political orientation) and this could even condition the emotional state of the respondent before the outset. Furthermore, it was a relief or escape following the concentration required in the main central part of the questionnaire.

(iii) Specific instructions for the procedure

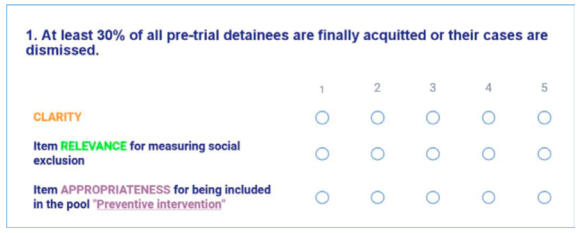
The purpose of the research and the theoretical concept were explained with two illustrative examples. In the sheet with the pools and procedural instructions there was an indication with regards to scoring the clarity, relevance and appropriateness of the items. In other words, this was intended to ensure compliance with the methodological process which is a suggestion "to elaborate on both the pools and the indicators being measured by each item of the test" (Escobar-Pérez & Cuervo-Martínez, 2008, p. 30).

As may be seen in the graph below, experts from the second group had to rate the clarity and relevance of each item on a double Likert scale (from 1 to 5). In this scale, designed on the basis of the Dunn protocol (Dunn *et al.,* 1999), 1 meant poor clarity or relevance of the indicator in explaining or measuring social exclusion, while 5 was the maximum score.



Graph 4. Questionnaire model of the second group.

The questionnaire model of the first group was very similar to that of the second group (see graph below), except for the inclusion of a third criterion: the appropriateness of the item to the predefined pools. Nevertheless, unlike the second group, these experts received, along with the questionnaire, an additional e-mail containing a circular letter explaining the progress of the project and the succeeding steps to be taken.



Graph 4. Questionnaire model of the first group.

(iv) Randomness of questions

The random arrangement of questions in the questionnaire prevents the "halo" or "contagion" effect —which results when questions whose answers may influence one another are placed together— from occurring (Sierra-Bravo, 2008, pp. 317-318). Upon remittance of the questionnaire to the second group, the classification of experts in one subgroup or another one was determined by the order of arrival of their name, thus ensuring randomness of experts in each questionnaire. However, questions were not distributed at random in each instrument, rather they were successively distributed by pools. This aspect was improved when the questionnaire was passed onto the first group: efforts were made not only so that personal and professional variables of the expert in each subgroup were as balanced as possible to achieve maximum extra-group homogeneity and intra-group heterogeneity, but also so that items of the partial questionnaires were distributed at random as well.

Regarding the way in which the questionnaire was managed, it was decided to implement it online or by e-mail for both groups of experts. All questionnaire models were created and sent by way of the *Qualtrics* platform, since it provides a professional and specialised service for this type of methodology. The questionnaire sent via e-mail offered many advantages (López-Roldan & Fachelli, 2015), and it was the most adequate method for our research, specifically: it is quick, inexpensive, allows simultaneous access to faraway places (a particularly useful aspect, given the geographical representation of the sample), avoids any influence stemming from the interviewer's action, enables carrying out a comprehensive follow-up of the questionnaire's completion status (open, on-going or finalised), measures the minutes spent in completing it, permits sending reminders, facilitates systematisation of the information received owing to its exportation to different databases and self-generates statistical reports.

In like manner, the *Qualtrics* platform, as opposed to other existing, less specialised programmes, had

its own e-mail server that sent electronic mails by using the e-mail addresses of the main researchers, even if they had been drawn up and supervised by other members of the coordinating team. This feature favoured constant communication before, during and after the completion of the questionnaire. Besides, this avoided distractions in the communication channel, because always the same interlocutors (the main researchers) sent and signed the messages, being closest to the expert and representing the greatest authority. Another advantage was that those cases rejected as spam were minimal, since the e-mail addresses were "trustworthy" or recognised by the e-mail server used by the experts.

Attributable to this method, there was a high response rate in each phase: 84.5% for experts of the second group (37% in part 1; 32% in part 2 and 31% in part 3) and 92.8% for those of the first group (in both parts of the questionnaire). The first group was the only one that carried out this task within the established period (two months), as opposed to the second group, that greatly exceeded it (they took approximately 15 months). As regards the time needed to complete the questionnaire, both groups took, on average, 20 to 30 minutes. Once this task was completed, a participation certificate —together with a gift— were sent to them by post and by electronic means.

Statistical analysis of results (Stage 3)

After each application session of the questionnaire, analyses were carried out from both a quantitative (scores) and a qualitative (experts' commentary) perspective. From a quantitative point of view, special attention was given to the results being drawn from the tests, since they were closely related to the content validity or, expressly, optimal statistical results ensured the validity and reliability of the items (Escobar-Pérez & Cuervo-Martínez, 2008). To this end, several tests were performed using basic statistical parameters (mean, mode, median, standard deviation, minimum and maximum ranges) in addition to inter-judge reliability tests in both groups, particularly Aiken's V validation coefficient (Aiken, 1980; 1985).

There are several ways to assess content validity by means of an agreement among judges. However, Aiken's V coefficient is the most adequate for its determination, as opposed to other similar approaches that are commonly used, such as the Index of Agreement (IA) and the Binomial Test (BT). Its main advantages are its simplicity of calculation (Merino-Soto & Livia-Segovia, 2009), its capacity to obtain values that can be compared according to the size of the sample of judges (Cordón, 2015), and the possibility of assessing its statistical significance (Escurra, 1988).

Aiken's V is regarded as:

> The ratio of a data obtained from the maximum addition in the difference of possible values. It may be calculated over the assessments of a group of judges with relation to an item or as the assessments of a judge with respect to 1. (Escurra, 1988, p. 107)

Its size includes a range of 0.00 up to 1.00 (Merino-Soto & Livia-Segovia, 2015). In this way, obtaining a 1 —which is the greatest possible magnitude— shows a perfect agreement among the evaluated judges and the smallest variability in their answers and, conversely, a low coefficient —lower than 0.50— means a lack of agreement among the judges. Formerly, the Aiken's V coefficient obtained was compared with a table of values (Cordón, 2015). Nevertheless, years later, Penfield y Miller (2004) introduced confidence intervals into the equation to determine the probability of occurrence of an event, taking into account the confidence level chosen by the researcher (Cordón, 2015). According to Merino-Soto and Livia-Segovia "as the size of the sample increases, the range of the interval shall be smaller and, therefore, calculating coefficient V will be more accurate" (2015, p. 171). In our case, the confidence interval was quite adequate (0.95), the validation coefficient cut-off was high (items with Aiken's $V = 0$ and/or > 0.70) and the sample size was very large compared to the one generally used for this methodology. All these aspects improved the test's accuracy.

By applying this statistical method to the second group of experts, 81 indicators —out of 126— with an Aiken's V higher than 0.70 were validated in the relevance criterion. This criterion was deemed predominant vs. the clarity criterion, because those items that were unclear could be reformulated with a more precise drafting in the second round of the questionnaire applied to the first group of experts. Out of the 81 validated items, 20 were validated with an Aiken's V in the 0.70 – 0.74 interval, 19 in the medium 0.75 - 0.79 interval and, nearly half of them (42), with an Aiken's V equal to or higher than 0.80. It should be noted that 62 of the 81 items validated for relevance were also validated for clarity and, particularly, 62 of them obtained 0.70 and over, 13 of them, 0.60-0.69, and 6 of them, 0.50-0.59. All in all, the results brought to light the high validity and reliability of the proposed indicators upon obtaining

64% of the total (81 of 126) with a coefficient greater than 0.70 in relevance.

A qualitative review was undertaken of the comments made by experts in the above-mentioned open field that was included in the questionnaire. Firstly, those comments that were just congratulatory of the work performed were discarded (2 comments) and the remainder (22 comments) focused on the following topics: confusion regarding the baseline theoretical definition, lack of clarity issues in the drafting of the item and its measurement, and difficulties understanding the task required of the expert. Regarding the first topic, it was considered that several indicators could lead to exclusion, but also to inclusion. As for the second topic, the comments manifested that some questions were long and difficult to understand; the negative wording used ("there is a lack of", "there is no", etc.) made the items less clear; certain terms could seem vague in different countries and their translation did not have the exact same ring as for the English-speaking individual. Some items seemed to measure two aspects in the same item, that is, they appeared to be double-barreled questions, which could affect their adequate measurement, as well as the fact of using certain percentages instead of others. Finally, in some cases it was not properly understood whether or not the expert's opinion on the items was required or if answers should reflect the reality of their country.

In view of the comments, we noted that a lack of clarity both in the wording of some indicators —due to their translation or complexity—, and in the explanation of the research objective and the procedure to be followed by the expert, was the main problem detected. In order to resolve it, three essential tasks were carried out: the reduction of the volume of validated indicators by reinforcing the eligible selection criteria, the re-writing those items that were unclear but could be included in the following phase and the sending a circular letter with a detailed description, and a reminder, of the research objectives as well as its progress. The first two tasks were implemented with the participation of the entire research group who, in successive meetings, established the following selection criteria: to maintain those items with 0.70 or higher scoring in clarity; to consider —as the case may be— those items with 0.70 or higher in relevance and over 0.54 in clarity; to remove items that were similar to others, i. e., that measure very similar aspects; to maintain a similar number of indicators in each pool; and to find a balance between rules and practices and among the pools themselves. Due to this process, the questionnaire was reduced from 81 to 65 items, that were distributed in nine pools (each with 5 to 11 items), 12 of which were rewritten in group sessions of the research team in order to improve their clarity. In short, a selection had been made of those items that were the most valid to be delivered to the first group of experts.

Final selection of indicators: the final instrument (Stage 4)

Once the questionnaire was completed by the first group of experts, we noted that validity as regards clarity increased in comparison with the previous round. As shown in the table below, on this occasion there was no indicator below 0.63, and therefore those items that had been rewritten no longer presented comprehension problems. Besides, the number of indicators in the intervals 0.75-0.79, and 0.80 and higher, increased considerably: they went from 41 items in the first round (25+16) to 57 (45+12). These results evidenced that the method used to clarify the most problematic items had produced good results. On the other hand, we noted a loss of validity in the relevance criterion, since 8 of the 65 items that had previously obtained a coefficient equal to or higher than 0.70 where now classified in the 0.63-0.69 interval. In other words, such items had obtained a larger consensus in the second group than in the first one. As regards appropriateness to the pool, there was only one item —with a coefficient of 0.35— that was not validated.

However, the 65 indicators obtained out of the final validation were still a very high number for achieving the management and dissemination of the instrument. Besides, 8 of them no longer passed the 0.70 cut-off limit. It was therefore decided to establish new methodological criteria, very similar to the ones established in the previous reduction of the instrument but, on this occasion, with two additional criteria: to obtain a number of items of around 35 but, in any case, they would not exceed 40; and to ensure a minimum of 3 items per pool (between rules and practices).

In order to carry out the screening, the research team analysed the first two items of each pool with the highest Aiken's V score in relevance, since predictably a greater agreement would be reached with regards to them, while validity increased again in the item's relevance, upon exclusion of those below 0.70. In the second round, the team reviewed the third item with highest scores and, in the third round, the remainder. Following this procedure, a final instrument of 39 items was achieved (see Díez-Ripollés & García-España, 2019), of which 26 were rules and 13 were practices. All had

Table 5.
Scoring of the 65-item-instrument according to the group and to criteria of clarity, relevance and appropriateness to the pool.

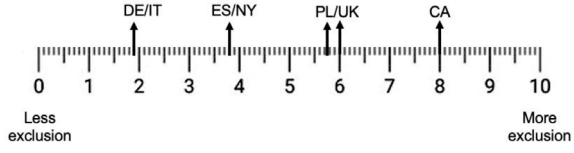
Aiken's V internals	Clarity (first group)	Clarity (second group)	Relevance (first group)	Relevance (second group)	Appropriateness to the pool
0.80 and higher	25	45	41	35	56
0.75 – 0.79	16	12	14	13	5
0.70 – 0.74	12	1	10	7	2
0.63 – 0.69	7	7	0	8	1
Not validated	5	0	0	0	1
TOTAL	**65**	**65**	**65**	**65**	**64**

coefficients that were equal to or higher than 0.70 in the three criteria and, in many cases, they exceeded 0.80 in clarity (31 items), relevance (30 items) and appropriateness to the pool (38 items). In summary, efforts were made to ensure that, by implementing a methodological procedure similar to the one adopted during the entire project to reduce the number of items, the final amount would consist of the most robust indicators for measuring social exclusion. The success of this methodological procedure was mostly due to the high number of items that were validated in both phases, given that such a high level of baseline validation provided flexibility to the research team at the time of choosing among them, as well as in establishing selection criteria that were ever more varied and demanding.

Ongoing development

Once the questionnaire had been validated, the research team took the first steps for its application. A pilot project was carried out to test its implementation in Spain (see García-España & Díez-Ripollés, 2021) collecting data regarding each indicator from different sources, using protocols created *ad hoc* and identifying possible biases associated with the criteria already employed. Subsequently, a purposive convenience sample consisting of 4 European countries and 2 US states (Germany, Italy, United Kingdom, Poland, California and New York[6]) was selected. We are currently collecting data for each of the 39 indicators in this sample, as we did in Spain[7]. The statistical procedure developed permits estimating the position of each country or state on a numerical exclusion scale ranging from 0 to 10, or from 0 to 100. Although it is still too early to know the final position of these territories on the scale, we can now observe that those considered exclusionary (USA and UK) in the international literature (Lappi-Seppälä, 2007; Nelken, 2009; Wacquant, 2000; Young, 1998) have obtained higher scores in contrast to Spain, Italy and Germany.

An example is the preliminary results obtained in pool 3 (Sentencing and sanctions systems), made up of items 33, 35, 36, 37 and 38, which highlight the potential of the instrument for measuring exclusion beyond the traditional indicators and their discriminatory capacity. For example, if we took into consideration only item 37, "The incarceration rate is higher than 120 inmates per 100,000 inhabitants", traditionally used to compare criminal policies, it would lead us to the conclusion that all countries are exclusionary except for Germany and Italy (the only ones under 120 prisoners). However, owing to the set of indicators that make up the dimension, the conclusion is quite different when estimating more precisely the level of exclusion for each country. In reality, California (CA) would be the most exclusionary state because it is closest to the pole of greatest exclusion, followed by the United Kingdom (UK) and Poland (PL).



Graph 5. Example of countries represented on a scale of exclusion ranging from 0 to 10 according to pool 3.

At the other end of the continuum, we find Germany (DE) and Italy (IT), the least exclusionary, followed at a certain distance by Spain (ES) and New York (NY) situated below 5 points.

6 The United States is being analysed according to states because we understand that each of them has individual characteristics. In this first application of the instrument, large states and from the East and West coast have been chosen.

7 As a part of the project: "Social exclusion as a criterion for criminal justice policy comparison: Application of the RIMES instrument. Reference: PGC2018-096073-B-I00, funded by the Ministry of Science and Innovation.

Methodological limitations

The great complexity and scale of the project did not lack methodological limitations —considered difficulties or aspects that could have been improved— throughout the process of creation and validation of the instrument. This section summarises the most important ones (see Tables 6 and 7) and a set of recommendations is proposed for future implementation, in order to substantiate this type of instruments.

During the creation phase of the instrument, the group of researchers that had been set up manifested an intermediate level of expertise, since several of its members did not consider themselves experts, either because they were in the early stages of academic and research training, or because they had not previously dealt with certain topics. Even though this circumstance was not restrictive, given that the pools were subsequently balanced with researchers of different levels, the replacement of two low-level members within the team of 20 experts would have increased the group's coefficient K. Thus, the Delphi method would have been applied in a more comprehensive way, since one of its objectives is the rejection by part of the coordinating team of those potential experts with worse self-assessments, and for that purpose, establish cut-off points for inclusion and for applying, usually, the Ebel method (Cruz, 2009).

As regards the consultation carried out to evaluate the clarity of the initial construct, the baseline was a theoretical concept already used by Díez-Ripollés in several publications and recognised in the academic field. In any case, it was decided to inquire of team members if they understood it properly and if they thought it was adequate for defining social exclusion. However, the request was sent by the coordinating team to the remainder of the group via e-mail, and this facilitated that —in spite of having requested just the opposite—, many researchers replied to the e-mail collectively, thus influencing one another and creating opinion subgroups. Further, in that request it was not specified that there were two aspects of the concept to be assessed: its clarity (whether it was understood or not) and its relevance in defining social exclusion. Even though the group sessions that took place successive to collecting the e-mails helped reach a consensus on the theoretical concept, they omitted the first step of individual reflection by means of a measurement scale for the objective quantification of the clarity and relevance of the concept. Likewise, no discussion or validation of the pools was carried out by the research team, which could have been undertaken with the procedure mentioned above, that is, by using scales.

In the instrument validation phase, the second and third stage also had limitations. It is well known that two of the main limitations of the questionnaire by e-mail method are the low rate of answers and slowness (López-Roldan & Fachelli, 2015). The first one did not take place since it was an intentional sampling, i. e., there was a mutual professional recognition among the parts (primary researcher and expert/s), which granted enough motivation to carry out the task. Besides, although some members of the second group were strangers to the coordinating team, the knowledge of the field of study and the large number of international experts participating was sufficient reinforcement of the commitment to the research. As regards the timeframe of the delivery of the instrument, there was a gap of one year between the period established for questionnaire application by the second group and its effective completion. Possibly the established deadlines were too optimistic, considering the large number of experts involved and the virtual nature of the reminder, which is less effective than a personal notification. Actually, this fact did not impact the quality of the

Table 6.

Limitations faced regarding the creation of the instrument.

Creation phases	Limitations	Recommendations
Phase 1. Setting up a group of researchers.	Partial development of the Delphi method. Not removing 2 low-level experts.	a) Full implementation of the Delphi method; and b) successive repetition of coefficient K´s test.
Phase 2. Consultation on the construct´s clarity.	Contagion of opinions on the concept among the research group´s members.	c) To combine scales of individual opinion (1 to 5) with an open field.
Phase 3. Operationalisation in pools.	The research team did not discuss or validate the pools.	d) To validate the pools with less clear (1 to 5) and an open field.
Phase 4. Selection of indicators.	None.	None.

Table 7.
Limitations faced regarding the validation of the instrument.

Validation stages	Limitations	Recommendations
Stage 1. Selection and contact of international experts.	None.	None.
Stage 2. Design and online management of instrument.	Slow answer of the second group. Difficulty to understand several items. No validation of the theorical concept and its pools. Lack of a pre-test with a reduced sample of experts.	e) To include validation of the theorical concept itself, as well as its explanation (in clarity and relevance) and of its pools. f) To better explain the terminology used. g) To include an additional column to score ease to answer the item. h) To include an additional column to score ease to answer the item. i) To adopt a more qualitative approach in the interviews with researchers. j) To conduct a pre-test.
Stage 3. Statical analysis of results.	Several items adversely affected by lack of clarity in the first phase of questionnaire management.	k) Several recommendations from the previous stage are applicable here.
Stage 4. Final selection of the strongest indicators.	None.	None.

research, it just delayed it. It was assumed that generating more reliable final statistical results was conditional upon obtaining a high answer rate.

Likewise, the quantitative and qualitative analyses of results from the second group, obtained in the first round of the questionnaire, evidenced difficulties in understanding several items, and even the theoretical concept and its pools. In this sense, it would have been positive to validate —both in clarity and relevance— the theoretical concept supplied and its pools, in order to determine whether the connecting link established between the theoretical sphere (definitions) and the practical one (indicators) was acceptable.

A glossary of specific terminology could have also been included, an additional column to determine the ease of answering the item (related to lack of clarity, inaccuracy of the wording or simple ignorance) and an open field for comments in each question. This latter improvement would have allowed the expert to indicate the most problematic items, while completing the questionnaire, and to explain the reasons wherefore. This information was lost as the final comment was too generic.

Some experts suggested that a more qualitative methodological approach based on interviews prior to the delivery of the questionnaire would have addressed many of the difficulties encountered. For example, if the questionnaire had been personally supplied to the experts that were native English speakers from the UK and/or the US, this would have reinforced the initial translation of the instrument. Likewise, if a review had been carried out to ascertain whether the instructions included in the questionnaire were sufficiently clear and precise, this would have reduced confusion as to the purpose of the research and the tasks to be carried out by the researcher. Actually, everything appeared to suggest the implementation of a pre-test or pilot study to obtain information on the operation of the instrument under real conditions. According to Fowler (1993) the best way to conduct a pre-test for a self-reported questionnaire is by using previous qualitative techniques. In our case, the pre-test was reduced to several technical tests in order to verify whether the questionnaire was properly sent and received, whether questions were properly registered or whether data could be downloaded, among others. If the guidelines suggested by Fowler had been applied, the instrument would have been completed in the first place by a reduced group of experts from different countries in the presence of the team coordinators, who would have led a joint discussion to clarify any problematic aspect.

All these difficulties were somehow reflected in the third stage, in view of several Aiken's V coefficients obtained regarding clarity. Even though afterwards such difficulties were remedied, in the subsequent round carried out with the first group of experts, it would have been convenient to follow the recommendations suggested for the second stage.

Conclusions

The methodology adopted in the two research projects that were developed has successfully responded to and fulfiled the main objectives of the study: to create and validate an instrument for criminal justice policy comparison that would enable classification of national crime control systems according to their socially exclusionary effects. Despite facing certain limitations —that were to be expected given the diversity of the methodologies used in the various

implementation phases over a long period of time—, a remarkably valid and reliable instrument for international comparison was obtained, which fills an important gap in the criminal justice policy knowledge.

The proposed set of 39 indicators measures, in a clear, comprehensive and discriminatory way, many aspects of criminal social exclusion, thus opening the door to analyses of comparative criminal justice policy that are more complex and useful than those traditionally focused on penal rigorism. Moreover, the good results obtained have proved the suitability of the content validation method by way of the judgment of experts for criminal justice policy research, as well as the relevance of Aiken's V test for the selection of robust indicators in large samples of experts. Likewise, the online management of the questionnaire did not present its traditional disadvantages.

As underscored by Díez-Ripollés y García-España (2019) the future implementation of this new tool in a high number of Western developed countries will permit comparison of their respective national penal justice policies, and they may be classified on a social exclusion scale. The results obtained from the effective implementation of the instrument in subsequent phases will certainly provide relevant data on the instrument's operation and the additional methodological challenges it faces.

In any case, it is expected that —in this advanced implementation phase—, the instrument can already become a useful tool to place the different national penal justice systems before a mirror, to promote substantial changes towards a criminal justice policy management that is less socially exclusionary.

References

Aiken, L. (1980). Content validity and reliability of single items or questionnaire. *Educational and Psychological Measurement, 40*, 955-959.

Aiken, L. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*, 131-142.

Blasco, J. E., López, A., & Mengual, S. (2010). Validación mediante método Delphi de un cuestionario para conocer las experiencias e interés hacia las actividades acuáticas con especial atención al windsurf. *Ágora para la Educación Física y el Deporte*, *12*(1), 75-96.

Cabero Almenara, J., & Barroso Osuna, J. (2013). La utilización del juicio experto para la evaluación de TIC: el coeficiente de competencia experta. *Bordón, 65*(2), 1-14.

Casas, F. (1989). *Técnicas de investigación social: los indicadores sociales y psicosociales (teoría y práctica)*. PPU.

Cordón, J. (2015). *Analizando la V de Aiken usando el método score con hojas de cálculo*. https://www.researchgate.net/publication/277556053_Analizando_la_V_de_Aiken_Usando_el_Metodo_Score_con_Hojas_de_Calculo

Crespo, T. (2007). *Respuestas a 16 preguntas sobre el empleo de expertos en la investigación pedagógica*. Editorial San Marcos.

Cruz, M. (2009). *El método Delphi en las investigaciones educacionales*. Editorial Academia.

Díez-Ripollés, J. L., & García-España, E. (2019). RIMES: An instrument to compare national criminal justice policies from the social exclusion dimension. *International E-Journal of Criminal Sciences, 1*(13), 1-27. https://www.ehu.eus/ojs/index.php/inecs/article/view/20866

Díez-Ripollés, J. L. (2011). La dimensión inclusión / exclusión social como guía de la política criminal comparada. *Revista Electrónica de Ciencia Penal y Criminología (RECPC), 13-12*, 1-36. http://criminet.ugr.es/recpc/13/recpc13-12.pdf

Díez-Ripollés, J. L. (2013). Social inclusion and comparative criminal justice policy. *Journal of Scandinavian Studies in Criminology and Crime Prevention, 14*, 62-78.

Ding, C. S., & Hershberger, S. L. (2002). Assessing content validity and content equivalence using structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 283-297.

Dunn, J. G., Bouffard, M., & Rogers, W. T. (1999). Assessing item content-relevance in sport psychology scale-construction research: Issues and recommendations. *Measurement in Physical Education and Exercise Science, 3*(1), 15-36.

Escobar-Pérez, J., & Cuervo-Martínez, A. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances en Medición, 6*, 27-37.

Escurra, L. (1988) Cuantificación de la validez de contenido por criterio de jueces. *Revista de Psicología*, *6*(1-2), 103-111. http://revistas.pucp.edu.pe/index.php/psicologia/article/view/4555

Estévez-García, J., & Pérez-García, M. J. (2007). *Sistema de indicadores para el diagnóstico y seguimiento de la educación superior en México*. Asociación Nacional de Universidades e Instituciones de Educación Superior.

Fowler, F. (1993). *Survey research methods. Applied social research methods series* (2nd ed.). Sage Publications.

Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings*. Kluwer Academic Publishers.

García-España, E., & Díez-Ripollés, J. L. (2021). La exclusión social generada por el sistema

penal español: aplicación del instrumento RIMES. *In-Dret, 1*, 259-281.

García, L., & Fernández, S. (2008). Procedimiento de aplicación del trabajo creativo en grupo de expertos. *Energética*, *XXIX* (2), 46-50.

González Blasco, P. (1994). Medir en las ciencias sociales. In M. García Ferrando *et al.* (comps.), *El análisis de la realidad social. Métodos y técnicas de investigación* (pp. 335-364). Alianza.

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing and Health*, *20*(3), 269-274.

Hernández-Sampieri, H., Fernández-Collado, C., & Baptista Lucio, P. (2006). *Metodología de la investigación* (4th ed.). McGraw-Hill.

Hyrkas, K., Appelqvist-Schmidlechner, K., & Oksa, L. (2003). Validating an instrument for clinical supervision using an expert panel. *International Journal of Nursing Studies*, *40*, 619-625.

Lappi-Seppälä, T. (2007). Penal policy in Scandinavia. In Tonry (ed.), *Crime, punishment and politics in comparative perspective. crime and justice* (n. 36). University of Chicago Press.

Lazarsfeld, P. (1985.) De los conceptos a los índices empírico. In R. Boudon & P. Lazarsfeld (ed.), *Metodología de las ciencias sociales*. Laia.

López, A. (2008). La moderación de la habilidad diagnóstico patológico desde el enfoque histórico cultural para la asignatura Patología Veterinaria. *Revista Pedagógica Universitaria*, *13*(5), 51-71.

López-Roldan, P., & Fachelli, S. (2015). *Metodología de la investigación social cuantitativa*. Universidad Autónoma de Cataluña.

Mehrens, W., & Lehmann, I. (1982). *Medición y evaluación en la educación y en la psicología*. Editorial CECSA.

Merino-Soto, C., & Livia-Segovia, J. (2009). Intervalos de confianza asimétricos para el índice la validez de contenido: un programa Visual Basic para la V de Aiken. *Anales de Psicología*, *25*(1), 169-171.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational measurement* (pp. 13-103). American Council on Education.

Nelken, D. (2009). Comparative criminal justice: Beyond ethnocentrism and relativism. *European Journal of Criminology*, *6*(4), 291-311.

Nunnally, J. (1973). *Introducción a la Medición Psicológica*. Buenos Aires: Paidós.

Oñate, M. (2001). *Utilización del método Delphy en la pronosticación: una experiencia inicial. Aplicación del Método Delphy*. www.rieoei.org/deloslectores/804Bravo.

Penfield, R. D., & Miller, J.M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement in Education*, *77*(4), 359-370.

Prieto, G., & Delgado, A. R. (2010). Fiabilidad y validez. *Papeles del psicólogo*, *31*(1), 67-74.

Rey del Castillo, P. (2004). Nota metodológica sobre los indicadores del barómetro del CIS. *REIS*, *108*, 151-178.

Sierra-Bravo, R. (2008). *Técnicas de investigación social: teoría y ejercicios* (14th ed.). Thomson Editores.

Skjong, R., & Wentworth, B. (2000). *Expert judgement and risk perception*. https://www.researchgate.net/publication/265032303_Expert_Judgment_and_Risk_Perception

Wacquant, L. (2000). *Las cárceles de la miseria*. Alianza Editorial.

Young, J. (1998). From Inclusive to Exclusive Society. In Ruggiero / South / Taylor (eds.), *The new European criminology. Crime and social order in Europe*. Routledge.