Journal of Criminal Justice 98 (2025) 102402

Contents lists available at ScienceDirect

Journal of Criminal Justice

journal homepage: www.elsevier.com/locate/jcrimjus

An item response theory approach to punitive attitudes

Nicolas Trajtenberg [a,*], Pablo Ezquerra [b,c]

[a] *Department of Criminology, School of Social Sciences, University of Manchester, Manchester, UK*
[b] *School of Social Sciences, Cardiff University, Cardiff, UK*
[c] *Facultad de Ciencias Sociales, Universidad de la República, Montevideo, Uruguay*

ARTICLE INFO

Keywords:
Punitive Attitudes
Item Response Theory
Graded Response Model
Psychometric evaluation
Scale reliability
Punishment
Public Opinion

ABSTRACT

Concerns about how punitive attitudes are measured are long-standing in the academic literature. However, empirical research in this area has often overlooked variations in the difficulty of punitive items and differences in individual dispositions. To address these challenges, this paper explores the measurement of punitive attitudes through the application of Item Response Theory to a representative sample of Uruguayan citizens. By addressing the limitations of Classical Test Theory in criminological research, we provide a more nuanced understanding of punitive attitudes, distinguishing the difficulty of survey items from respondents’ underlying traits. Our analysis highlights significant variation in item difficulty and discrimination, showing that assuming that all punitive measures reflect equivalent levels of punitiveness is problematic. Our findings identify gaps in the measurement of individuals with higher levels of punitiveness and suggest that frequently used survey items may fail to capture the full spectrum of punitive attitudes. This research emphasizes the need to refine survey instruments to enhance the validity and reliability of scales of punitive attitudes, thereby contributing to a more comprehensive understanding of public opinion regarding crime and punishment.

1. Introduction

By 2019, Uruguay had seen a sustained rise in violent crimes over previous decades, fuelling public concern about security and demands for tougher policies (Sanjurjo et al., 2021). That year’s presidential elections included a plebiscite proposing tough-on-crime measures such as incorporating the Armed Forces into a National Guard, eliminating early release for certain serious offences, introducing revocable life imprisonment, and authorising night raids. Although 47% of voters supported the proposal, it fell short of the majority needed for approval. These results raised questions about public attitudes toward crime and punishment. Critics argued that the reform was too punitive, especially with the inclusion of measures like life imprisonment (de Sierra, 2019; Llabrés & Torres, 2020). Did the outcome suggest most Uruguayans were not punitive? The ’vote’s aggregated nature—combining multiple proposals—complicated the interpretation of ’citizens’ attitudes. During those years, polling firms conducted surveys to assess the potential public support for punitive measures. For example, 83% approved prohibiting early release for serious crimes, 79% endorsed introducing life imprisonment, and 43% supported the death penalty for exceptionally serious crimes (Opción Consultores, 2018; Subrayado, 2019). These figures raised questions regarding what percentage of Uruguayans could be considered to have ’punitive ’preferences’ and potentially vote for those reforms. Is endorsing at least one of these punitive policies sufficient to categorise someone as punitive? How should we interpret those who support two measures—are they twice as punitive as those who only support just one? Furthermore, should all measures carry equal weight, or is endorsing life imprisonment less punitive than supporting the death penalty? These questions reveal the complexity and multifaceted nature of attitudes and the challenges in defining and measuring them through surveys. They highlight the need for nuanced methodologies to better understand public opinion on punishment-related issues.

Concerns about how we conceptualise ’citizens’ punitive attitudes – understood as a preference for punishing offenders and increasing the costs of crime (Ramirez, 2013) – are not new in the academic literature (Adriaenssen & Aertsen, 2015; Aguilar-Jurado, 2018). Research has shown that public support for tougher punishments can significantly influence the justice system, often resulting in harsher sentencing practices and higher incarceration rates (Enns, 2014; Jennings et al., 2017). Moreover, such punitive approaches that usually focus on extended prison terms and widespread incarceration may not effectively

[*] Corresponding author at: Department of Criminology, School of Social Sciences, University of Manchester, 176 Oxford Rd, Manchester M13 9QQ, UK.
E-mail address: nico.trajtenberg@manchester.ac.uk (N. Trajtenberg).

https://doi.org/10.1016/j.jcrimjus.2025.102402
Received 20 January 2025; Received in revised form 26 March 2025; Accepted 26 March 2025
Available online 10 April 2025

reduce crime. Instead, they can produce several negative side effects, such as higher rates of reoffending, human rights violations, and significant financial burdens (Cullen et al., 2011; Liu et al., 2018; Loeffler & Nagin, 2022).

In recent years, empirical research has increasingly focused on understanding the measurement challenges associated with this complex construct (Ezquerra, 2025). Initial approaches relied on broad, one-dimensional indicators, such as support for the death penalty or a general preference for harsher sentencing, that often lacked contextual nuance. These measures may have inflated perceptions of punitiveness by failing to account for alternative responses and by ignoring key details such as the nature of the crime, the offender, or the victim (Kruis et al., 2023; Stalans, 2013). Over time, more sophisticated methods have emerged. Some surveys introduced vignette-based questions, presenting respondents with specific case scenarios that allowed for the control and manipulation of crime characteristics in experimental settings (Campregher & Jeglic, 2016; Horstman et al., 2021; Socia et al., 2019). Others have employed multi-item scales validated through psychometric techniques like exploratory and confirmatory factor analysis to improve measurement reliability (Aizpurúa, 2015; Armborst, 2017; Maguire & Johnson, 2015; Trajtenberg et al., 2024).

Along with the introduction of question sets designed to better capture the latent dimension to be measured, researchers have increasingly focused their analysis on its dimensionality and components, significantly influencing the items included in surveys. Adriaenssen and Aertsen (2015), in their review of the literature, identify four main dimensions associated with the construct of punitive attitudes: 1) the goals of punishment, associating retributive, deterrent, and incapacitative goals with punitive attitudes, as opposed to rehabilitative or restorative goals; 2) the forms or types of criminal punishment, associating certain types of punishment such as the death penalty or life imprisonment with punitive attitudes as opposed to other types of punishment that do not involve prison, like community service or house arrest; 3) the intensity of punishment, associating longer sentences with punitive attitudes; and finally 4) sentencing policies, such as 'three strikes and you are out' laws versus more discretional sentencing policies. However, the presence of these elements in the batteries of questions used in the literature has been inconsistent. Moreover, the extent to which these dimensions encompass the entire concept and the extent to which they differ from each other is still under debate (Armborst, 2017; Maguire & Johnson, 2015).

Despite these advancements, the empirical literature is based on the assumption that all survey questions are equally relevant or revealing of 'citizens' punitive attitudes. It does not carefully consider the difficulty or intensity of each question, or how it connects to different dimensions of punitive attitudes – including not only the most direct aspect, 'severity of penal sentences', but also the broader punitive and progressive justifications of punishment (e.g. rehabilitation, retribution, deterrence, etc.). Additionally, it does not consider that individuals might have different dispositions related to their punitive attitudes that affect how they respond to questions.

To address these challenges, we applied an Item Response Theory (IRT) approach to a survey on punitive attitudes conducted with a representative sample of Uruguayan citizens in 2018. Our article begins by reviewing the differences between Classic Test Theory (CTT) and IRT and the advantages of the latter. It also reviews the literature, highlighting the scarce application of CCT in criminological topics, particularly regarding attitudes toward punishment. Next, we describe our methodological approach, including our sample, measures, and analytical approach. Last, we present our main findings contextualising them in the literature on punitive attitudes and considering how they might inform future lines of research.

1.1. Classical test theory

Despite considerable progress in psychometric assessment and modelling over the past decades, its incorporation into criminological testing practices is limited not only to punitive attitudes but also to most theoretical constructs. While in other fields like psychology, education or health, the use of IRT has led to major improvements in test development and validation (der Linden & Wim, 2017; Reise & Waller, 2009), in criminology, its use remains limited to the use of CCT approach based on Reliability or Factor Analysis (Piquero et al., 2002; Sweeten, 2012). Although there is some criminological research based on IRT on some specific areas or particular constructs, such as self-control (Gibson et al., 2010), control balance theory (Piquero et al., 2001), offending scales (Johnson & Raudenbush, 2006), or even risk assessment tools for the prediction of recidivism (Giguère & Bourassa, 2023), most of the studies use analysis techniques based on CCT analysis (e.g. Factor Analysis, Cronbach alpha). This type of analysis faces some relevant challenges in the context of measurement. First, under the CCT approach, estimate scores are highly sensitive to the sample of respondents and the specific items used in the survey (Hambleton et al., 1991). In other words, there is no specific consideration of differences in respondents' abilities, and all items are assumed to contribute equally to the total score. Additionally, given that the CCT approach is more focused on the total score of the construct, it is more limited when it comes to the analysis of individual items and the identification of those that are biased or less informative (Cai et al., 2016), and provide less detail on how precision may vary on different levels (Nguyen et al., 2014). Finally, under the CCT theory approach, the equivalence of items across different groups or samples is assumed but never actually assessed (Nguyen et al., 2014; Wilson, 2023). More advanced approaches like Item Response Theory (IRT) are used to address some of these limitations, as they provide a more detailed understanding of the measurement process and item functioning.

1.2. Importance of item response theory

Measuring punitive attitudes is not a simple and straightforward task. As with most psychological constructs, they are abstract, not directly observable, and lack single items or empirical referents (Cronbach & Meehl, 1955). Thus, the measurement approach assumes that the construct we want to measure is latent and is defined by indicators in a specific way: empirical items or indicators are caused by or reflections of the construct (Blalock, 1968; Bollen & Lennox, 1991).

Item Response Theory (IRT) is a latent variable model based on this measurement approach. It aims to show how this underlying latent trait affects individuals' responses to multiple empirical items related to a scale that measures a single construct (Baker, 2001). Nevertheless, it presents some relevant differences and advantages with regard to classical test theory. There are some key assumptions for the model to produce valid and interpretable results that include: (i) items of the scale reflect a single underlying trait or construct (*unidimensionality*); (ii) each item's response is influenced by the latent trait and not by other items (*local independence*), (iii) as the level of the latent trait increases, it also increases the probability of having a positive or higher response to each item (*monoticity*); and (iv) item characteristics are stable across different groups of respondents (Nguyen et al., 2014; Wilson, 2023).

Imagine Nico and Pablo taking a survey with 10 questions about their punitive attitudes, and they have to answer whether they endorse statements regarding punishment and criminal justice institutions, scoring one point for each affirmative answer. If both get a total score of five points, is it safe to conclude that both are equally punitive? Probably not. Maybe Pablo has got his points for endorsing more severe items such as '*Death penalty for recidivists is a fair* '*punishment*' while Nico is getting his points for items such as '*individuals that commit crimes deserve to be punished because they have harmed* '*society*'. Respondents are less likely to endorse items if they are more severe or difficult. Additionally, does the score of Nico and Pablo's punitive attitudes depend only on the severity level of the items they are answering? To answer affirmatively, we must assume that Nico and Pablo have identical underlying traits of punitiveness. This is a strong assumption and usually does not hold. If some individuals have a stronger underlying trait regarding punitive attitudes, they are more likely to endorse punitive items. These two ideas regarding the heterogeneity of individual traits and the severity of items, as well as how to disentangle them, lie at the heart of the IRT approach.

One of the main disadvantages of the CCT framework is that it assumes that estimates of scores will be identical irrespective of attribute/ latent trait levels of respondents, which leads to '*sample* '*dependent*' test scores, that is, scores are highly sensitive to the distribution of traits of the individuals in the sample (Hambleton et al., 1991; Zanon et al., 2016). Thus, the proportion of respondents rated as punitive will vary significantly depending on the characteristics of the sample. For example, samples composed disproportionately of males, older individuals, or people of low socioeconomic status who have experienced victimisation or trust less in the criminal justice system (Applegate et al., 2002; Davey et al., 2024; Dodd, 2018; Payne et al., 2004; Spiranovic et al., 2012) are more likely to show a higher proportion of punitive respondents compared to more heterogeneous samples. Thus, punitive items in the former samples may inaccurately appear more difficult or severe than punitive items in the latter samples (Zanon et al., 2016). A second related problem is that test global scores are '*test dependent*', that is, if items used in the survey are less difficult or severe, global scores will be low and vice versa, even if their underlying 'ability' remains constant (Hambleton et al., 1991). Thus, individuals might score higher and look more punitive if they have to answer a survey with less severe or easier items, and more progressive if more extreme or intense items were included. In other words, items are treated as being completely interchangeable, with little idea about their different weights in terms of difficulty, but also their potential problems of overlapping or redundancy in terms of the specific construct (Gordon, 2015). These two types of item and sample dependencies pose a serious obstacle to creating and evaluating items and scales that can adequately compare individuals based on their level of punitiveness.

The IRT is a latent trait model that offers a solution to these problems by estimating the probability of responding to items that depend on both item characteristics and respondents' characteristics (aka, their latent trait) (Baker, 2001). This approach allows estimating the severity/difficulty of items independently of specific samples and how the underlying trait is distributed across its respondents in each sample (Piquero et al., 2001). Moreover, we can measure a person's ability independently of the specific choice of items (Lord, 2012). Under IRT, unidimensionality involves that individuals' performance on items is caused only by this unobserved latent trait, and thus, once we control for it, there are no additional dependencies between items (Hambleton et al., 1991). Thus, how an individual responds to one punitiveness item would not affect their response to another once we consider their punitiveness trait.

One of the most interesting aspects of IRT is that it is better suited to assess the weight or severity of items on a scale. Items are no longer equivalent and interchangeable; instead, they are situated in different places along a continuum of intensity or severity, which is a function of an individual's underlying trait or ability (Cai et al., 2016; Gordon, 2015). In this sense, one key aspect is the *item characteristic curves (ICC)*, which is a probability curve that models the relationship between individual latent trait (theta) and their response to each specific item on a scale (Hambleton et al., 1991). Each ICC is represented in a graph that plots the probability of endorsing the item (or the response category in the case of polytomous items) on the Y-axis and the respondents' underlying latent trait on the X-axis. ICC allows the visualisation of item key parameters. First, the *difficulty or location parameter* shows the location or level of the underlying trait where the item has a 50% probability of being endorsed (Baker, 2001). The higher the parameter, the more severe or difficult the item. Respondents will be more likely to respond yes (or choose the higher option of a Likert scale) to 'easier' items situated lower in the continuum and involving less severity of punitive attitudes (e.g.' *individuals that commit crimes should receive some penal punishment*'). Likewise, those harder or more severe items located higher in the continuum (e.g.' *recidivists should be punished with the death penalty, no matter the nature of the crime*') will be less likely to be endorsed and would indicate higher degrees of punitiveness. Second, the discrimination parameter reflects how effectively an item differentiates between individuals with varying levels of the trait: the steeper the curve, the higher the discrimination, meaning that minor differences in the individual trait will lead to larger differences in the probability of endorsing the item than those less discriminating items (Baker, 2001). Additionally, for polytomous questions, ICCs allow for the assessment of whether response categories are adequately functioning in terms of their relative ordinality and the extent to which all categories are effectively used, and they provide meaningful information at different levels of the trait for each item.

Additionally, ICC and *Function Curves* provide relevant information for a more targeted selection and elimination of those least informative items based on their very low level of discrimination or extreme difficulty levels (Dillard et al., 2013). Information Function Curves allow examining with more detail the precision measurement that each item contributes and how it might vary across different levels of the underlying construct, both at the item level (*Item Information Curves*, IIF) and by the whole scale level (with a *Test Information Function Curve*, TIF) (Cappelleri et al., 2014).

Another tool of the IRT approach is the *Wright Map* or *item person map,* which plots in the same scale the distribution of items according to the level of difficulty (the most difficult ones at the top and the easiest ones at the bottom), and individuals according to their latent traits (the individuals with highest ability at the top and with lowest ability at the bottom). The Wright Map allows assessing whether the scale and its item difficulties are adequately distributed across all the levels of the underlying construct and are aligned to measure the range of abilities of individuals in the sample (Wilson, 2023). It can reveal measurement gaps or lack of comprehensiveness when there is an uneven distribution of items across the underlying construct, resulting in scarce or null items measuring certain areas (e.g., extreme levels of difficulty or certain middle ranges) (Gordon, 2015). It can also reveal if there is a mismatch between item difficulties and individual abilities. For example, the Wright map might show three different types of problems: locations with persons but no corresponding items, indicating the scale is failing to measure those specific levels of person ability; empty spaces that are populated by items but with no or few persons, indicating either that items are too easy or too difficult to endorse; or finally locations of item redundancy, where too many items with similar level of difficulty are concentrated at particular points of the latent trait, assessing the same ability level without adding unique information (Baker, 2001; Piquero et al., 2000; Piquero et al., 2001).

Infit and *outfit* statistics are other diagnostic tools used to identify inconsistent and problematic response patterns. While Infit identifies unusual values from respondents whose ability level is close to the 'items' difficulty level on the underlying latent trait, Outfit detects unexpected responses from respondents with abilities positioned distant from items' location on the underlying trait continuum (Linacre, 2002). Basically, higher infit/outfit values indicate '*noisy items*' that show inconsistent patterns of responses, with lower-ability respondents endorsing difficult items and vice versa, signalling potential problems of wording or misinterpretation (Wind, 2016). Instead, lower values indicate '*uninformative*' or '*mute*' *items,* which are those that are too predictable and are answered in the same way by all the sample. Thus, they add little unique information to the precision of the construct because they are not challenging or differentiating among individuals with heterogeneous abilities (Wind, 2016).

Finally, an additional feature is the evaluation of measurement equivalence at the item level. *Differential Item Functioning (DIF)* models item–trait interaction to evaluate the degree of item fairness or consistency, that is, whether individuals with similar levels of abilities of the underlying trait, but from different groups (e.g. gender, age groups, ethnic or cultural backgrounds), have a different likelihood of endorsing items (Wilson, 2023). For example, if males and females with similar punitive attitudes show a different probability of endorsing a specific item (e.g., recidivists should be punished with 'prison'), we say that this item is biased. In such cases, observed differences might reflect measurement bias rather than true differences in punitive attitudes between sexes.

1.3. Item response theory in criminology

The IRT approach has been used in criminology in the last two decades in three main areas. First, when it comes to *predictor or explanatory constructs*, research has focused on the assessment of Grasmick et al. (1993) operationalisation of Gottfredson and Hirschi (1990) construct of self-control in youth populations, with particular attention to the validity or bias of items, category endorsement, as well as differential functioning across different groups of gender and self-control (Gibson et al., 2010; Higgins, 2007; Pechorro et al., 2023; Piquero et al., 2000; Rocque et al., 2013; Ward et al., 2010). Beyond self-control, some interesting exceptions are the assessment of a cynicism scale among police officers showing if categories are used adequately, bias in some items, and lack of equivalence across genders and white and non-white respondents (Hickman et al., 2004); the evaluation of Tittle's (Grasmick et al., 1993; Tittle, 2018) control balance ratio scales which also focuses on misfit items of the construct, how categories are endorsed and the existence of gender bias in terms of item functioning (Piquero et al., 2001); or how the incorporation of IRT improves the construction and validation of procedural justice scales (Graham et al., 2025). Second, some authors have also used IRT to address limitations in *criminologists' outcome variables*, particularly traditional summative offending scales. For example, using response patterns to multiple self-report crime/delinquency items to generate latent traits to identify more relevant/severe and less biased items, as well as evaluate if 'items' are invariant across different groups (e.g. age, sex, race, residence) (Conrad et al., 2010; Johnson & Raudenbush, 2006; Osgood, Finken, & McMorris, 2002; Osgood, McMorris, & Potenza, 2002; Piquero et al., 2002). Finally, there are some applications for the psychometric evaluation of *risk assessment tools,* such as the Level Service/Case Management Inventory (LS/CMI), identifying redundant or scarcely discriminant items and their effect on predictive validity and reliability, as well as differences in how items function across gender (Giguère & Bourassa, 2023; Giguère, Bourassa, & Brouillette-Alarie, 2023; Giguère, Higgs, & Charette, 2023).

However, when it comes to punishment and punitive attitudes, as far as we know, all the empirical research has been focused on the psychometric evaluation of dimensionality based on CTT using Cronbach's alpha, Exploratory and Confirmatory Factor Analysis (Aizpurúa, 2015; Armborst, 2017; Maguire & Johnson, 2015; Mascini & Houtman, 2006; Ortet-Fabregat & Pérez, 1992; Ramirez, 2015; Silver & Silver, 2017; Trajtenberg et al., 2024). Thus, despite improvements in the last decade in the psychometric assessment of attitudes toward penal punishment, we believe that applying an IRT approach would offer several benefits. First, by disentangling difficulty and ability, this analytical approach allows the assessment of items that differentiate respondents across various levels of punitiveness, identifying gaps where current scales may lack coverage. Second, a more specific analysis of item characteristics and functioning (e.g., discrimination, information) is a valuable tool to better understand which items help to capture the punitive attitudes construct and minimise measurement error. Additionally, the reliability and validity of punitive scales are enhanced by ensuring that each item's contribution is not identical but weighted according to its discrimination power. Finally, the use of sample-independent metrics improves the generalizability and comparability of psychometric findings across different populations and cultural contexts.

1.4. Present study

In this paper, we use a large representative sample of the population in Uruguay, where respondents completed self-report questions regarding their attitudes toward penal punishment. We applied the IRT approach to answer the following questions:

Under the 'unidimensionality assumption' that punitive attitudes constitute a single construct, we seek to answer the following exploratory questions:

- Are there any items showing low levels of discrimination among respondents or significant misfits that suggest they do not add much information to the global scale?
- Is there a bias in items regarding sex, age or social class?
- Are there significant differences in terms of difficulty in punitive attitudes usually used interchangeably?
- Is the scale of punitive attitudes adequately distinguishing individuals with different levels of ability and items with different difficulty levels, or are there significant measurement gaps?

2. Methods

2.1. Participants and measures

The data (n = 895; males = 430; females = 465) was collected through a representative telephone survey in Uruguay using a sample of cellular phones and a Computer – Assisted Telephone Interviewing in a project on Trust in Criminal Justice System and Punitive Attitudes funded by the Open Society (see specific details on sampling, weights and adjustment in Trajtenberg et al., 2024).

The measure used is a punitive attitudes scale consisting of 14 items, which recent psychometric assessments have shown to be empirically unidimensional (Trajtenberg et al., 2024). Items included are in line with the main themes of the punitive attitudes literature. First, we include six items measuring the intensity or severity of the punishment including items that refer to general increase of severity of sentences '(*'Make sentences more severe for all* '*crimes*'), use of death penalty for serious crimes (*'person convicted of murder should receive the death penalty*'), or expressing preference for being harsher with young offenders ('*Juveniles who commit serious crimes should be treated like adults*'). Second, we included items that capture different punitive reasons for punishing involving both deontological or just deserts justifications such as retribution ('*Criminals deserve to be punished because they have harmed society*') or more teleological ones such as incapacitating or isolating offenders to prevent further victimization ('*We should put criminals in jail so that innocent citizens will be protected from criminals who will victimise them*') or deterring offenders by showing that committing a crime is very costly ('*Punishing criminals more harshly would reduce crime by setting an example and showing others that crime does not pay*'). Finally, the progressive rationale for punishment in the criminal justice system was incorporated in three items that consider both rehabilitation or treatment of offenders ('*The rehabilitation of prisoners has proven to be a failure*') and restorative justice principles ('*Confronting perpetrators with the sufferings of their victims prevents them from relapsing*'). Responses for each of the 14 items were recorded on a Likert-type scale, ranging from 1 (totally disagree) to 7 (totally agree) (see appendix for description of all items).

In order to carry out the Differential Functioning Item, we included the variable sex coded as male = 0, female = 1; the variable race, originally coded as: black = 1, Asian = 2, white = 3, indigenous = 4, mestizo or mulatto = 5, other = 6 and recoded as non-white = 0 and white = 1 (due to the low number of cases in the non-white categories); and a seven-category ordinal variable of socio-economic status based on 10 questions from a socio-economic index that capture key aspects such as income, housing conditions and access to services. For the comparison between the IRT index and the frequentist index, we included respondents' educational level with levels from 1 = no formal education to 14 = college degree; religious orientation originally coded as 1 = catholic, 2 = evangelic, 3 = Jewish, 4 = none, and 5 = other, recoded as 0 = non-religious, 1 = any religion; and finally if they have suffered victimization in the last 12 months coded as 0 = none, 1 = yes.

2.2. Data analysis

Samejima (1968) Graded Response Model was used to assess the punitiveness scale for the Uruguayan sample, given that it is the most appropriate for polytomous models with ordered categories where discrimination is not constant across different items (Nguyen et al., 2014). The Graded Response Model estimates the probability for an individual to select each of the seven response categories for the punitive attitudes' items given his/her latent trait of punitiveness and thus, distinguishes among respondents with higher and lower levels of punitive attitudes. Assessment of model parameters followed Baker and Kim's (2017) guidelines, indicating that values below .65 are considered a low level of discrimination, values above 1.34 are considered high, and values over 1.69 are considered very high. Additionally, the Wright map was plotted to assess the distribution of items and persons across the underlying trait of punitive attitudes and evaluate if there are relevant measurement voids or if the scale encompasses a sufficiently broad range of respondents with different levels of punitive attitudes as well as items with different levels of intensity that cover the entire continuum of the underlying latent trait. Additionally, infit/outfit statistics were estimated for each item to identify items that are either below .5 or above 1.5 and thus are either not very informative or show inconsistent patterns that degrade the scale (Linacre, 2002). Finally, potential bias or lack of equivalence between males and females was evaluated using Differential Item Functioning. Analyses were performed using the R Studio program (R Core Team, 2021), particularly the packages mirt (Chalmers, 2012), lordif (Choi et al., 2011), and WrightMap (Torres Irribarra & Freund, 2014).

3. Results

3.1. Parameter estimation

IRT analysis was conducted for the 14 items as a composite measure of punitive attitudes, given the support for the unidimensionality of punitive attitudes in recent research (Aizpurúa, 2015; Armborst, 2017; Trajtenberg et al., 2024).[1] The parameter estimates and fit statistics for all items are presented in *Table 1*.[2] Discrimination parameters, that is, how well each item distinguishes between individuals with different levels of punitiveness, varied significantly, with the restoration item (p24) showing the lowest score (.88) and one of the severity items referring to juveniles (p51) showing the highest value (2.59) (see also Fig. 3). None of the scale items showed low levels of discrimination (below the .64 threshold), but items p24 and rehabilitation being a failure item (p59) show flat information curves (Fig. 6, Appendix), suggesting they provide less information and are ineffective in discriminating punitiveness between respondents across different construct levels. On the other extreme of the spectrum, 50% of the scale items are highly effective at distinguishing between respondents with -with different levels of punitiveness, with the two severity items referring to juveniles (p50 and (p51) as the most discriminant of the scale (*Fig. 3*).

Differential Item Functioning (DIF) analysis indicated that the punitiveness scale functioned similarly across sex and race, but somewhat less consistently across socio-economic status groups. This suggests that observed differences in item endorsements largely reflect true differences in punitive attitudes, rather than measurement error or bias. No item showed equivalence problems between white and non-white respondents, and only one severity item related to too much concern with criminal rights (p19) lacked equivalence between males and females. Yet, there is more reason for concern regarding socioeconomic status as three severity items related to the death penalty, relaxing controls on police to improve crime, and treating juveniles as adults (p16, p18 and p50 respectively), as well as an incapacitation item advocating to jail offenders and throw the key (p61) and, showed measurement bias (*Table 1*).

3.2. Item functioning

We assessed the fit of each item in the punitiveness scale using outfit/infit analyses. None of the 14 items showed serious misfits either as too noisy or erratic (higher than 1.5 threshold) or too weak to discriminate the full range of response levels (lower than .5 threshold). Still, six items show infit/outfit outside the ideal range (below .85 or above 1.2), and thus, while they are still productive, they can be considered degrading or generate more distortion in the scale, introducing noise or reducing overall precision. This degrading aspect of items is particularly salient for the items stating that rehabilitation programs allow criminals to get away with crime (p20) and the item about increasing the severity of all sentences (p15) (*Fig. 1*).

The use of categories is problematic across items on the scale. Not only did all items fail to utilise all seven response categories effectively, but there was also a significant overlap in the probability of choosing categories at their maximum probability threshold. In practical terms, 10 items are frequently used as either no/yes dichotomous items, where respondents either disagree completely or agree completely with the item, or at best, a third category is used, but with a very low probability of endorsement (*Fig. 2*). Nevertheless, the use of the lowest and highest categories is ordered: the higher the respondents' level of punitiveness of the respondent increases, the chances of selecting the higher category of total agreement with the item also increase and vice versa.

The person–item or Wright map in *Fig. 4* displays the distribution of i) persons reported levels of punitiveness, ii) and of items according to their level of difficulty or intensity. Respondents with lower scores in their punitive attitudes are at the bottom of the histogram, while (and less high punitive attitudes are at the bottom top. Similarly, while items that measure higher levels of punitiveness (i.e. more difficult items) are located toward the top of the diagram, items that measure lower levels of punitiveness (easier to endorse) are at the bottom.

Fig. 4 suggests a relatively normal distribution of persons and items spread across a significant part of the latent construct of punitive attitudes ranging from -1.96 to .17. Thus, the scale does a good job of distinguishing between individuals with varying levels of punitive attitudes. Yet, while the distribution of persons is more spread —particularly toward the top, where those with more punitive attitudes are located —punitive items are less spread and slightly more

[1] Support for the unidimensionality of punitive attitudes in Trajtenberg et al. (2024) is based on multiple analyses. EFA using the Kaiser criterion and Velicer's MAP supported a one-factor solution, although parallel analysis suggested a six-factor solution. In CFA, the single-factor model showed acceptable fit (CFI = .982, TLI = .979) but had a χ^2/df ratio of 5.23 and an RMSEA of .079, both exceeding recommended thresholds. A three-factor model marginally improved CFI (.983) and TLI (.980), yet the χ^2/df ratio (5.06) and RMSEA (.078) remained high, and a Heywood case indicated difficulty in distinguishing subdimensions. In contrast, a bifactor model incorporating one general factor and two subordinate factors achieved better overall fit (χ^2/df = 4.24, RMSEA = .06, CFI = .992, TLI = .989) and yielded an explained common variance (ECV) of .855, with most items loading more strongly on the general factor. Reliability estimates (omega = .85, hierarchical omega = .80) further support the unidimensionality hypothesis, as confirmed by an exploratory bifactor analysis.

[2] Although the Graded Response Model was selected for theoretical reasons, the Rasch/Partial Credit Model as a less complex alternative, showed a worse fit as indicated by higher AIC, BIC, and SABIC values, a lower log-likelihood, and a significant likelihood ratio test ($\chi^2(13) = 782.61$, $p < .001$) (see Table IV in the appendix).

Table 1
Graded Response Model Item Parameters

		Category thresholds						DIF		
	Discrimination (a)	*b1*	*b2*	*b3*	*b4*	*b5*	*b6*	Sex	SES	Race
p15	2.07	-1.93	-1.65	-1.34	-1.04	-0.5	-0.18	0.011	0.497	-
p16	1.56	-0.53	-0.31	-0.06	0.24	0.7	0.95	0.891	0.000*	-
p18	1.46	-0.96	-0.8	-0.54	-0.35	-0.01	0.33	0.037	0.000*	-
p19	1.61	-2.01	-1.77	-1.43	-1.07	-0.58	-0.23	*0.005**	0.125	-
p20	1.68	-1.75	-1.41	-1.01	-0.55	-0.08	0.29	0.872	0.641	-
p21	2.11	-1.66	-1.35	-1.09	-0.79	-0.4	-0.01	0.580	0.921	-
p22	1.54	-3.06	-2.8	-2.29	-1.8	-1.2	-0.61	0.110	0.209	-
p23	1.8	-2.07	-1.76	-1.49	-1.15	-0.64	-0.26	0.704	0.399	-
p24	0.88	-1.84	-1.29	-0.77	-0.13	0.82	1.36	0.167	0.630	-
p25	1.91	-2.51	-2.32	-1.8	-1.36	-0.86	-0.39	0.309	0.897	-
p50	2.19	-1.28	-0.97	-0.72	-0.44	-0.08	0.23	0.289	0.038*	-
p51	2.59	-1.52	-1.26	-1.02	-0.72	-0.33	-0.01	0.573	0.937	-
p59	1.02	-2.94	-2.38	-1.79	-1.02	-0.32	0.34	0.935	0.324	-
p61	1.89	-1.1	-0.75	-0.46	-0.11	0.3	0.58	0.647	0.000*	-

Note: a = estimated slope discrimination parameters (measures how well the item discriminates between respondents with different levels of punitiveness); b1, b2, b3 and b4 = estimated category threshold parameters (represent cut-off points for different response categories); DIF = Differential Functioning Item (indicates if item shows measurement invariance across groups) (threshold alpha = .01)

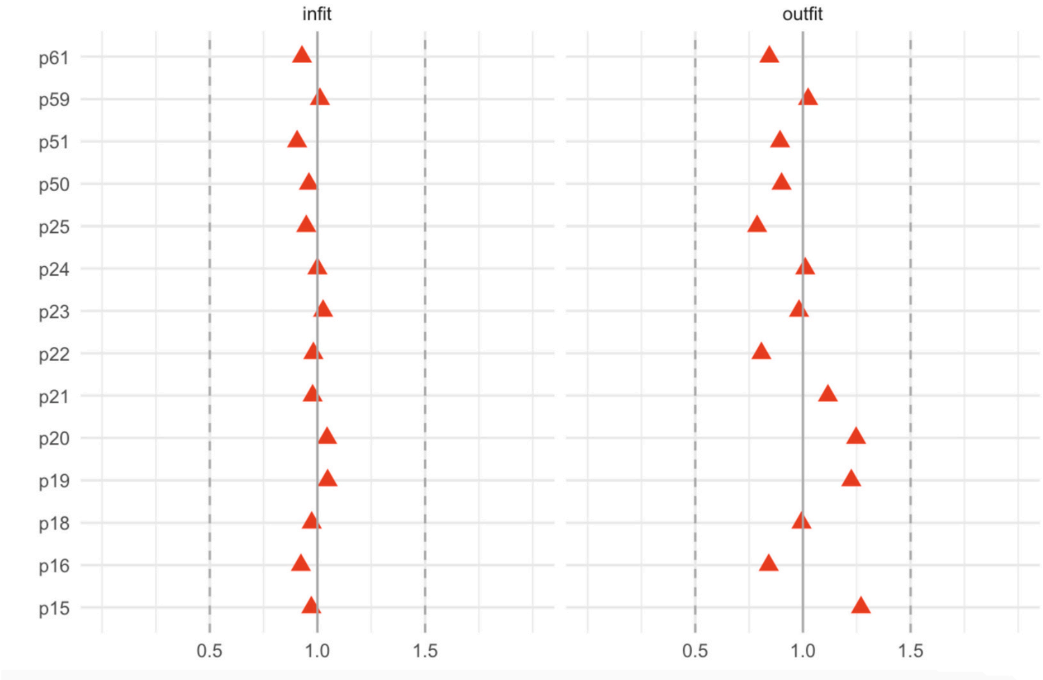

Fig. 1. Item Infit and Outfit statistics for each item. Note: Infit values indicate how well items function for respondents whose ability is near items' difficulty level, while outfit values detect unexpected responses from respondents whose ability is far from the item's difficulty level. Values outside range 0.5 – 1.5 suggest potential misfit

concentrated toward the bottom, suggesting many items are easier to endorse by individuals with lower levels of the latent trait-punitiveness. Additionally, the mean of persons and items scores do not match: the mean of the punitiveness score of persons is close to nearly one standard deviation (.78) below the mean of item difficulty. In other words, on average, punitive items are relatively too easy for a considerable part of the sample of persons (see *Fig. 7* in the appendix), particularly the item tapping on retribution (p22) and one of the items tapping on incapacitation to protect future victims (p25), which are the easiest to endorse. Conversely, the harsher incapacitation item that refers to 'throw away the key' (p61) and the severity item that refers to death penalty (p16) are the most difficult items, which measure better the most punitive persons.

Our findings in Fig. 4 also show *no gaps with items with no corresponding persons*. All items effectively capture relevant points along the punitive attitudes construct continuum, with no item measuring extreme levels of the trait with no respondents. Although the items capture most individuals of the sample, there is a relevant *gap with persons but no items*. That is, the assessment gap seems concentrated in the higher extreme suggesting that the punitiveness scale may not fully capture the most punitive respondents in the sample. Finally, there is no significant problem of redundancy, and items contribute uniquely to different difficulty levels.

The Scale Information Curve (SIC) in *Fig. 5* shows that punitive attitudes are not homogeneously measured across different levels of the trait. The peak information, reflecting the highest precision of the scale, occurs between -1 and 0 levels on the θ scale. While the maximum values of the Test Information Curves demonstrate that the scale has more than acceptable level of precision, with peak values above 15, the asymmetrical and relatively narrow shape suggests reduced precision and

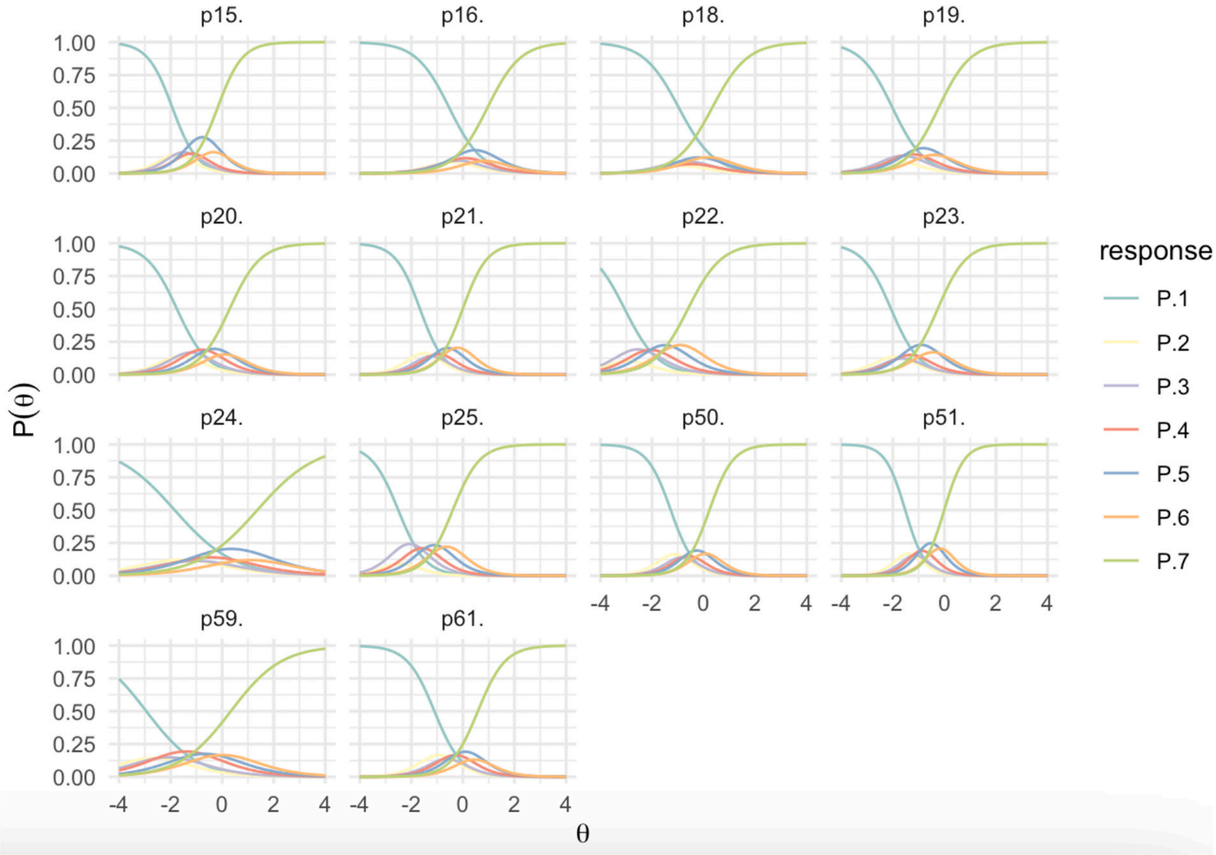

Fig. 2. Item Characteristic Curves (ICC) for items of the punitiveness scale. Note: ICC shows the relationship between respondents' underlying trait levels (punitive attitudes) and their probability of endorsing specific response categories for each item.

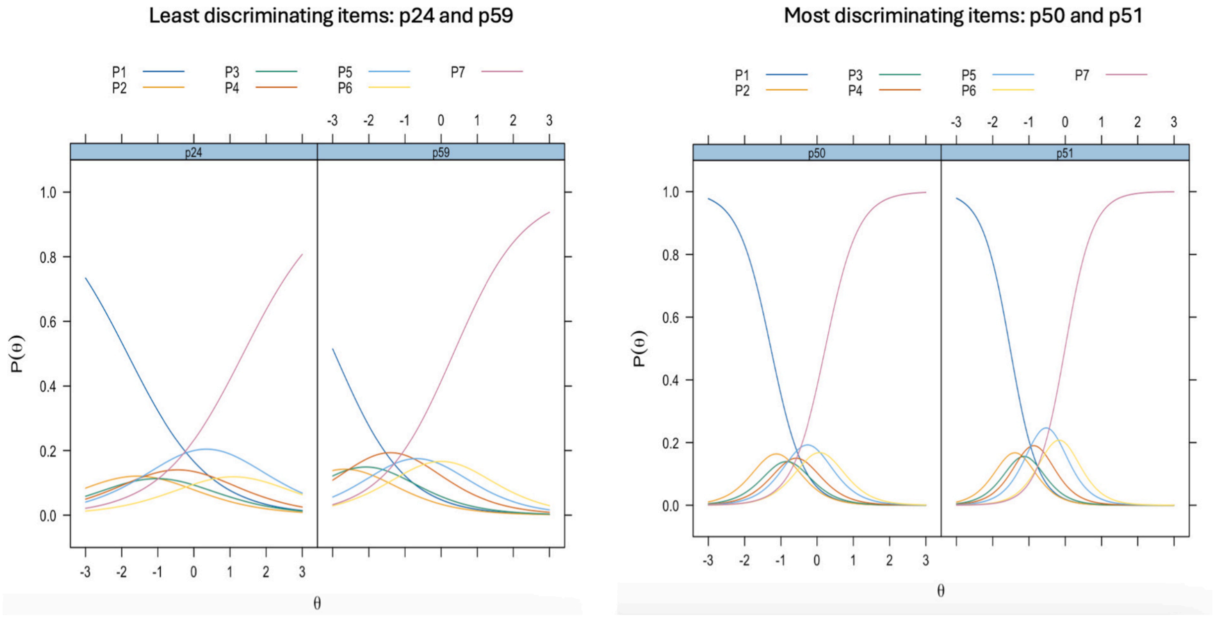

Fig. 3. Illustration of ICC curve for the items with highest and lowest discrimination ability

information at both extremes, especially at the higher end of the scale between 1 and 4 values of θ.

The comparison of the IRT index and a simple frequentist index shows a high and significant correlation, as expected (spearman rho = .97). However, Fig. 6 shows that for lower values of the simple summatory index, there appears to be a certain degree of underestimation of punitive attitudes compared to the inferences made using IRT. In contrast, for higher values of the simple summatory index, the reverse happens leading to an overestimation of punitive attitudes. For instance, the two green points highlight the potential divergences between both

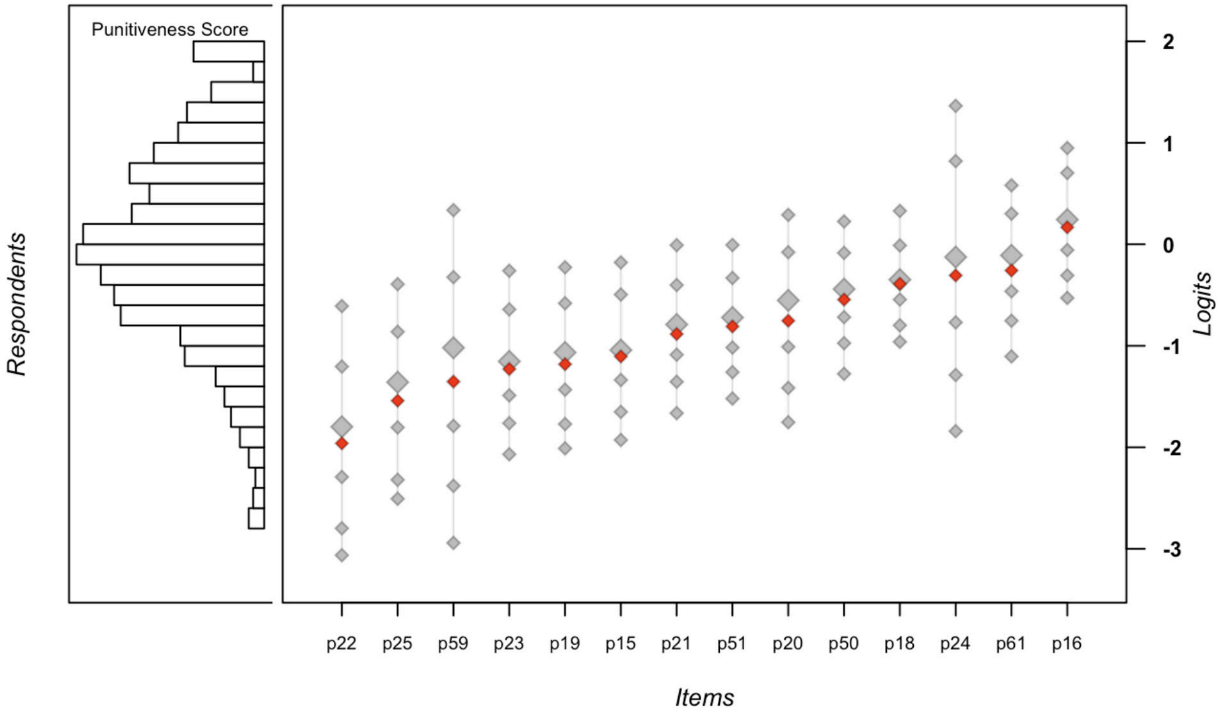

Fig. 4. Wright map of Punitive Attitudes Scale. Note: The Wright Map shows the relationship between respondents' punitive attitudes latent trait (left) with individuals and item intensity (right) on a shared scale. Higher values at the top indicate stronger punitive attitudes for respondents and more intense items

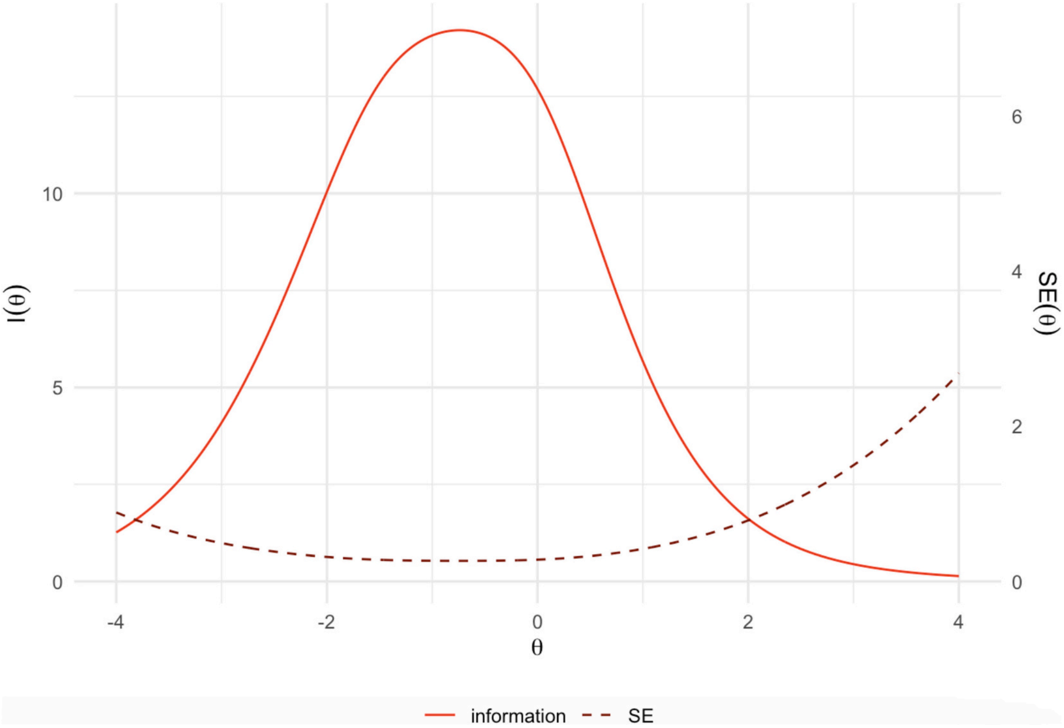

Fig. 5. Test information Curve of punitive attitudes scale. Note: The Test Information Curve shows how well the punitive attitudes scale measures different levels of the latent trait, with the x-axis representing punitive attitude levels, the y-axis representing measurement precision, and peaks indicating where the scale most reliably differentiates punitive attitudes among individuals

indexes. These two cases illustrate a situation where, according to the simple summatory index, the left case is ranked lower than the right case (and thus is less punitive), whereas the opposite occurs when comparing these cases based on the theta values extracted from IRT. Additionally, IRT index allows for a more refined estimation, increasing the variance in the measure of punitive attitudes, as evidenced by the dispersion around the regression line marked in red. Finally, when comparing both indexes after standardization, although both show similar relations with specific correlates of punitiveness such as education, victimization or religious orientation, the size of the coefficients differs (see Table V in

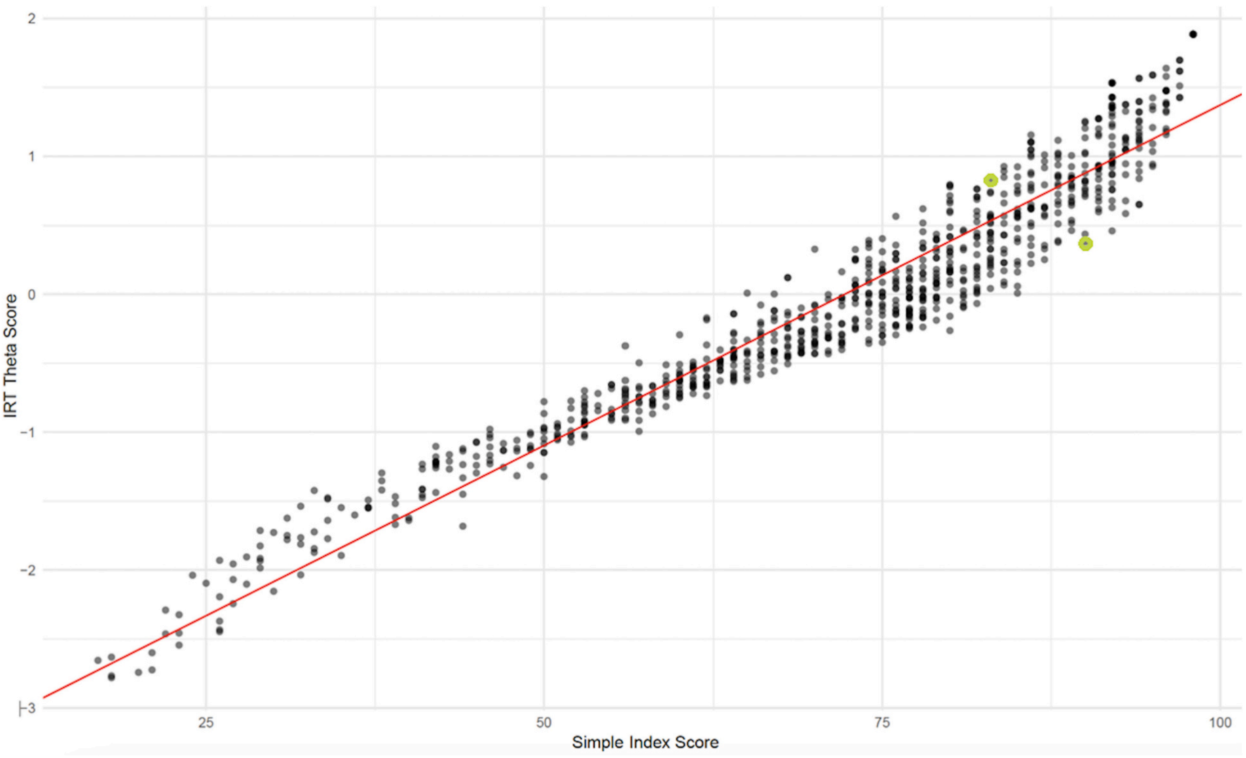

Fig. 6. Comparison between IRT Index and Frequentist Index

the appendix), suggesting that some finer distinctions may be overlooked when we choose a simpler frequentist approach.

4. Discussion

Latent constructs play an important role in criminological discipline and thus need a rigorous empirical examination (Mathesius & Lussier, 2021; Rocque et al., 2013). While in some topics such as self-control (Pechorro et al., 2023), cynicism (Hickman et al., 2004) or control balance (Piquero et al., 2001), criminologists have gone beyond exploratory and confirmatory analysis, when it comes to attitudes toward punishment research remains underdeveloped. If we want to understand why people have punitive attitudes and how they can promote ineffective or even iatrogenic policies (Enns, 2014; Jennings et al., 2017; Nivette, 2016), we must take its measurement seriously. To fill this gap, we undertook the first application of IRT to this construct. Our paper provides evidence that IRT results can be helpful for test development and improve the reliability and validity of punitive attitudes scales.

One of the most relevant conclusions of our study is that the assumption that items can be considered exchangeable, ignoring difficulty levels and weights, does not hold. Our results show that items have very different levels of difficulty, with some being particularly difficult to endorse (e.g. preference for the death penalty or indefinite incapacitation for chronic recidivists), while others are very easy to endorse (e.g. need for punishment when harming society). Previous research has included items on punitiveness as if they all have equivalent difficulty levels (Davey et al., 2024; Metcalfe & Cann, 2020; Simmler et al., 2021). At best, some studies seek to obtain balance by including items with more progressive content (e.g. regarding rehabilitation or human rights) and more punitive content (e.g. focus on the increase of penal costs or reference to retribution or incapacitation goals) or by using methods that such as CFA to balance the estimation of the latent trait but without explicitly modelling item difficultly (e.g. Armborst, 2017; Intravia, 2019; Trajtenberg et al., 2024). However, this balance in terms of content and methods does not guarantee that different levels of difficulty of items are considered or even that there is no redundancy of items. Additionally, our results also provide evidence that some items that tap on the failure of rehabilitation (p59) or concern of the criminal justice system with offender rights (p24) and are used in the literature (Cullen et al., 1985; Tam et al., 2008; Viney et al., 1982) are not particularly informative and might be worth reconsidering their inclusion (or reformulation) when measuring punitive attitudes. Furthermore, our results indicate a note of caution regarding the misfit of items. Some items, such as increasing severity for all crimes (p15), courts' preference for the rights of criminals over victims (p19), rehabilitation programs to escape punishment (p20), and prison as a general solution to protect citizens (p25) which are standard practice in scales in criminological studies (Brand & Anastasio, 2006; Cullen et al., 1985; Cullen et al., 1988; Hogan et al., 2005; Spiranovic et al., 2012; Viney et al., 1982) show degrading levels of misfit and are worth considering for their exclusion or reformulation by future studies.

Many of these item problems not only might undermine the reliability and validity of the construct itself but, if left unaddressed, might introduce biases in studies that use these constructs for predictive or explanatory purposes. Punitive constructs generated without accounting for variance in item difficulty run the risk of over or underemphasising specific latent trait ranges, which might lead to bias in factor scores (Zanon et al., 2016). Similarly, including items that show misfit or add little information and precision to the estimation of the latent construct can decrease the reliability and validity of factor scores (Nguyen et al., 2014). Thus, several criminological studies that include punitive attitudes as their main dependent variable, either in regression models (e.g., Costelloe et al., 2018; Davey et al., 2024; Intravia, 2019) or structural equation modelling (e.g., Armborst, 2017; Gerber & Jackson, 2016), run the risk of, at best, yielding partial or inaccurate estimates that attenuate true relationships with correlates by introducing measurement error, and at worst, inflating spurious associations, leading to biased or false positive results in predictive models.

One of the most interesting findings from the Wright map is the relative mismatch between respondents' levels of punitive attitudes (persons' abilities) and the difficulty of items designed to measure them. While our pool of items captures the lower levels of the punitive attitudes construct fairly well, it falls short in addressing the higher end of the spectrum. This finding indicates that our scale fails to adequately measure a significant subset of individuals with stronger and more severe views on punishment, highlighting a broader limitation in how punitive attitudes are often assessed. Test Information Curve confirms this problem, showing that our scale is better designed for individuals with less extreme punitive attitudes and becomes less reliable and imprecise when evaluating respondents with stronger punitive views. This issue is relevant in punitive attitudes research, where traditional survey items often overestimate these attitudes because they lack contextual information (Stalans, 2013). Vignette-based surveys, by contrast, provide specific crime scenarios, reducing respondents' reliance on exaggerated mental images of crimes and offenders that can bias judgments of penal severity (Kruis et al., 2023; Socia et al., 2019). However, our findings suggest that the issue is more complex and even traditional surveys may fail to capture the higher end of punitive attitudes, potentially underrepresenting the most punitive segment of respondents. This result is puzzling given that our scale includes extreme items like support for the death penalty (p16), indeterminate sentences expressed in a very extreme form ('lock them and throw the key) (p61), or even some items taping on harsher punishment on juvenile offenders (p50 and p51). Future research should test if the more punitive individuals can be more adequately measured using new items that tap into using severe punishments such as the death penalty or indeterminate sentence for minor crimes (Durham & Alexis, 1988); or that focus on severe penal sanctions of vulnerable minorities (e.g. migrants, ethnic minorities, or individuals with mental health issues) (Armborst, 2017; Guillermo et al., 2021); or including more expressive and retributive motivations such as support for public shaming, corporal punishment or even torture (Conrad et al., 2018; Slyke et al., 2018).

Our study provides evidence about how items are placed on the scale, and particularly regarding the alignment of item difficulty with *sanction severity* rather than with *punishment goals*. The two items associated with more severe punishments, such as the death penalty (p16) and indeterminate sentencing (p61), are located at the more difficult end of the scale, which aligns with more extreme punitive attitudes. Conversely, items that tap into less severe sanctions or are vaguer—merely suggesting that punishment is needed—are placed at the easier end of the scale. This evidence suggests that, in contrast with previous studies (Courtright et al., 2005; Falco & Martin, 2012; Gault & Sabini, 2000; Mackey & Courtright, 2000), the severity of the punishment, rather than the underlying rationale of punishment, plays a more significant role capturing the extremes of the distribution of the construct of punitiveness. For instance, retributive intuition tends to be the easiest to endorse, which is consistent with literature suggesting that most people think about the law in retributive terms (Carlsmith, 2008; Giacomantonio & Pierro, 2014; Keller et al., 2010). However, other aspects, such as anti-rehabilitation views, are also placed at the easier end and show little variation. Items related to incapacitation, instead, are also spread across both extremes of the scale, with the more difficult one emphasising severity (e.g. life sentences for repeat offenders), while the easier ones make more vague references to penal severity (e.g. imprisonment for dangerous criminals to protect citizens). All in all, these findings raise important questions about how we think about conceptualizing and measuring punitive attitudes. Future assessment of punitive attitudes constructs could benefit by focusing on the intensity of punishment and less on the underlying philosophical justifications (e.g., retribution, deterrence, rehabilitation, etc.). These goals of punishment could then be used as predictor variables rather than defining the difficulty of items (Ostaszewski et al., 2024; Ezquerra, 2025).

In criminology, the study of punitive attitudes has followed a trajectory similar to that described by Rozin for social psychology and Goertz for political science, prioritising causal modelling and hypothesis testing over descriptive work. Much like social psychology, where Rozin (2001) critiques the lack of extensive examination of phenomena before advancing formal and statistical models, criminology has often sought to identify key determinants and explain punitive attitudes without first adequately conducting a rigorous description of them. Similarly, Goertz's (2006) argument in political science that description should not merely work as an instrument for causal theories, but a core scientific task on its own applies here: the criminological field has largely overlooked the independent value of detailed measurement and exploration of the punitive attitudes construct. By failing to develop robust measurement, criminologists risk drawing conclusions from poorly defined or inconsistently operationalised constructs, undermining the validity of subsequent predictive models. Still, it is relevant to consider the research goals, number of items, sample size, simplicity, and transparency of the method since previous evidence from psychological and educational studies shows that IRT approach is not the default winner option in every context (Fan, 1998; Jabrayilov et al., 2016). Future research using psychometric approaches, such as Item Response Theory, can help bridge this gap by providing a more nuanced and rigorous understanding and validation of punitive attitudes constructs – provided it considers carefully the trade-offs with CTT approaches.

This study includes some limitations which open opportunities for future research by scholars of punitive attitudes. First, while IRT allows researchers to estimate item properties independently of specific samples, applying it across different cultural contexts presents challenges. Attitudes toward punishment are shaped by cultural, legal, and societal contexts, which can vary significantly across regions and jurisdictions (Kornhauser, 2015; Mayhew & van Kesteren, 2002). It is important to explore whether certain items might be universally 'difficult' or less likely to be endorsed (e.g., support for the death penalty or rejection of impunity) or universally easy or more likely to be accepted (e.g., prioritising harm reduction), and which might be culturally idiosyncratic. Likewise, are the different reasons for punishment ranked similarly in terms of difficulty across societies? For instance, the difficulty of endorsing incapacitation, and particularly the use of imprisonment, may be influenced by how commonly prison sanctions are used in a given society (e.g. imprisonment rates). However, a critical methodological challenge is determining whether variations in responses to punitive attitudes across societies reflect true differences in cultural norms, policies, or legislation—or are artefacts of bias in item interpretation. The systematic use of DIF analysis in future cross-national studies will be instrumental for future cross-national research to identify sources of variation and ensure the validity of punitive attitude measures across diverse contexts.

Second, more research is needed to explore whether differential item functioning across subgroups affects the validity of our results beyond cross-cultural comparisons. This study showed no significant threats to measurement invariance between males and females and between white and non-white, though 4 out of 14 items displayed significant DIF across socio-economic status groups, suggesting some variability in measurement. However, this analysis should be replicated with samples that allow for more nuanced comparisons among race groups beyond the white/non-white distinction. Additionally, future research should assess other characteristics that have been associated with attitudes toward punishment and offenders and could generate bias in the construct, such as economic anxieties (Singer et al., 2020), age (Payne et al., 2004), previous victimisation (Simmler et al., 2021), fear of crime (Armborst, 2017) or trust in authorities (Gerber & Jackson, 2016).

Finally, another limitation is the specific set of items evaluated in our study. While we included several items by relevant authors and studies in the field of punitiveness, the final pool of items has limited content diversity. Specifically, it lacks sufficient representation of both punitive and less punitive reasons for punishment, as well as a variety of punishment targets (e.g., migrants or marginalised populations). Future research should replicate this analysis with alternative items that capture a broader spectrum of restorative and deterrence-oriented perspectives or focus on the punishment of marginalised populations and evaluate whether novel items better address the measurement gaps. Additionally, response categories used for punitive items require further inspection. While our results indicate that responses generally function as practically dichotomous, more research is needed to test whether 4- or 5-point Likert scale responses might provide more nuanced measures of punitive attitudes.

5. Conclusions

This study highlights the importance of applying more sophisticated measurement approaches, such as IRT, in the study of punitive attitudes. Our findings show that current punitive attitude scales fail to fully capture the variability and complexity of citizen attitudes, particularly at the higher end of the punitive spectrum. While the scale effectively differentiates moderate or low punitiveness, they miss nuances among those with extreme views, potentially underestimating the prevalence of strongly punitive perspectives. By identifying poorly fitting items and low discrimination power, this research highlights the importance of refining scales and developing new items that better capture the full spectrum of punitive attitudes. Additionally, IRT helps detect item bias, ensuring more equitable measurement instruments. Although no significant bias by gender was found, future research should explore potential biases related to factors like socioeconomic status, ethnicity or previous victimization to contribute to more generalizable and representative findings. Overall, this study advocates for a paradigm shift in the measurement of punitive attitudes, a necessary precondition for understanding the social and political factors driving punitiveness and preventing the reinforcement of an unfair and inefficient criminal justice system.

CRediT authorship contribution statement

Nicolas Trajtenberg: Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Pablo Ezquerra:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation.

Funding

This work was supported by the Open Society Foundation Grant awarded to the project: 'Punitiveness in Uruguay: how punitive are Uruguayans and what are the main reasons?'. The funders had no role in the study design, data collection and analysis, interpolation of these data, decision to publish or preparation of the manuscript.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

We thank the two reviewers for their constructive feedback and helpful suggestions to the orig

Appendix A. Appendix

Table II
Item Descriptions, Dimensions, and Descriptive Statistics

Items	Label	Dimension	Mean (std. dev)
p15	Make sentences more severe for all crimes	Severity	5.65(1.9)
p16	A person convicted of murder should receive the death penalty	Severity	3.72(2.48)
p18	It is all right for the police to break the law in order to better control violent crimes	Severity	4.54(2.54)
p19	In general, our courts have been more concerned with the rights of criminals than victims	Severity	4.42(2.17)
p20	All rehabilitation programmes have done is to allow criminals who deserve to be punished to get off easily	Progressive goals (rehabilitation)*	5.02(2.12)
p21	Punishing criminals more harshly would reduce crime by setting an example and showing others that crime does not pay	Punitive goals (general deterrence)	5.37(2.08)
p22	Criminals deserve to be punished because they have harmed society	Punitive goals (retribution)	6.22(1.39)
p23	The amount of punishment that a criminal receives should be equal to the harm that the victim of the crime was forced to suffer	Punitive goals (retribution)	5.71(1.88)
p24	Confronting perpetrators with the sufferings of their victims prevents them from relapsing	Progressive goals (restoration)	4.24(2.23)
p25	We should put criminals in jail so that innocent citizens will be protected from criminals who will victimise them	Punitive goals (incapacitation)	6.05(1.54)
p50	Juveniles who commit serious crimes should be treated like adults	Severity	4.9(2.28)
p51	Make sentences more severe for juveniles who commit crimes	Severity	5.39(2.07)
p59	The rehabilitation of prisoners has proven to be a failure	Progressive goals (*rehabilitation*)*	5.33(1.91)
p61	Since most criminals will commit crimes over and over again, the only way to protect society is to put these criminals in jail and throw away the key	Punitive goals (*incapacitation*)	4.4(2.35)

[*] Inverted

Table III
Factor loadings of items

	F1	h2
p15	0.773	0.598
p16	0.676	0.457
p18	0.651	0.424
p19	0.687	0.472

(*continued on next page*)

Table III (*continued*)

	F1	h2
p20	0.702	0.493
p21	0.779	0.606
p22	0.671	0.45
p23	0.727	0.529
p24	0.458	0.21
p25	0.747	0.558
p50	0.789	0.623
p51	0.836	0.699
p59	0.513	0.263
p61	0.743	0.552

Table IV
Model Fit Comparison: Rasch/Partial Credit Model vs. Graded Response Model

	AIC	SABIC	HQ	BIC	Loglik	X2	df	p
Rasch/partial credit model	35843.71	35981.50	35999.51	36251.44	-17836.86			
Graded Response Model	35087.11	35245.96	35266.73	35557.19	-17445.55	782.606	13	.000

Table V
Comparison of Regression Estimates for the IRT Index and Simple Frequentist Index Across Education, Victimiazation and consumption of news

Variable	Dependent variable	Estimate	Std. Error	T value	Pr(>|t|)
Education	Frequentist index	-0.05177	0.006789	-7.625	6.22e-14
Education	IRT index	-0.04987	0.006805	-7.329	5.21e-13
Victimization	Frequentist index	0.13453	0.08421	1.598	0.11
Victimization	IRT index	0.12483	0.08422	1.482	0.139
Religious orientation	Frequentist index	0.30445	0.06615	4.602	4.78e-06
Religious orientation	IRT index	0.27840	0.06628	4.200	2.93e-05

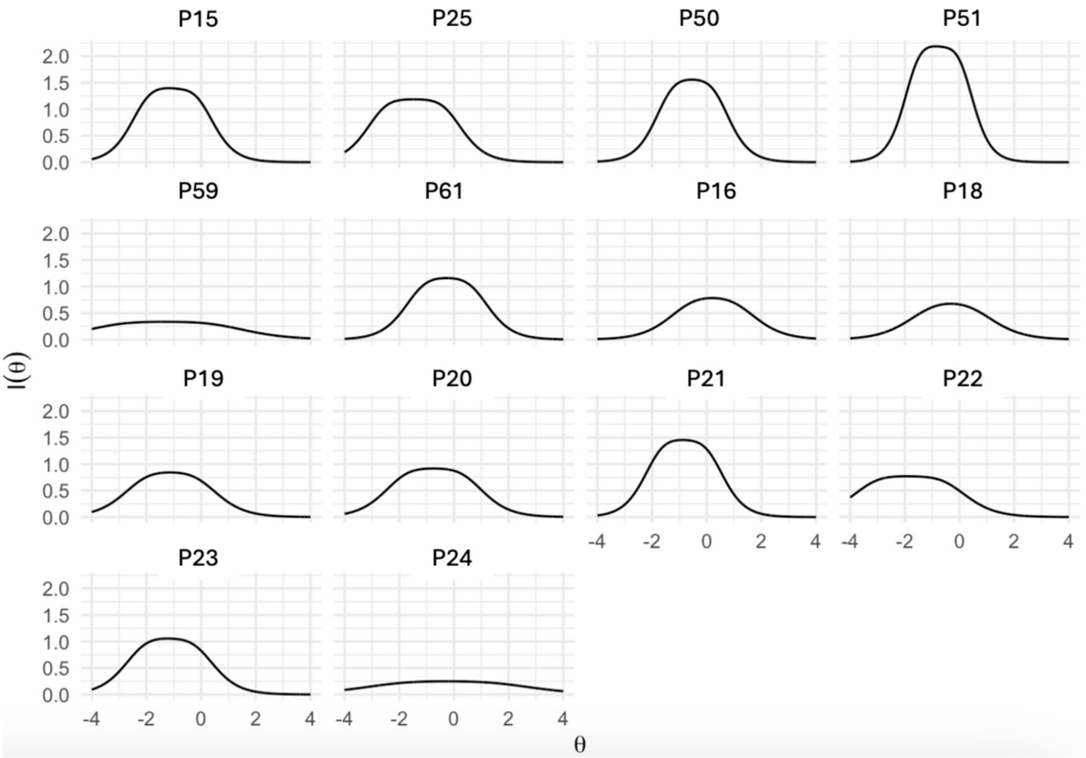

Fig. VI. Test Information Curves of items

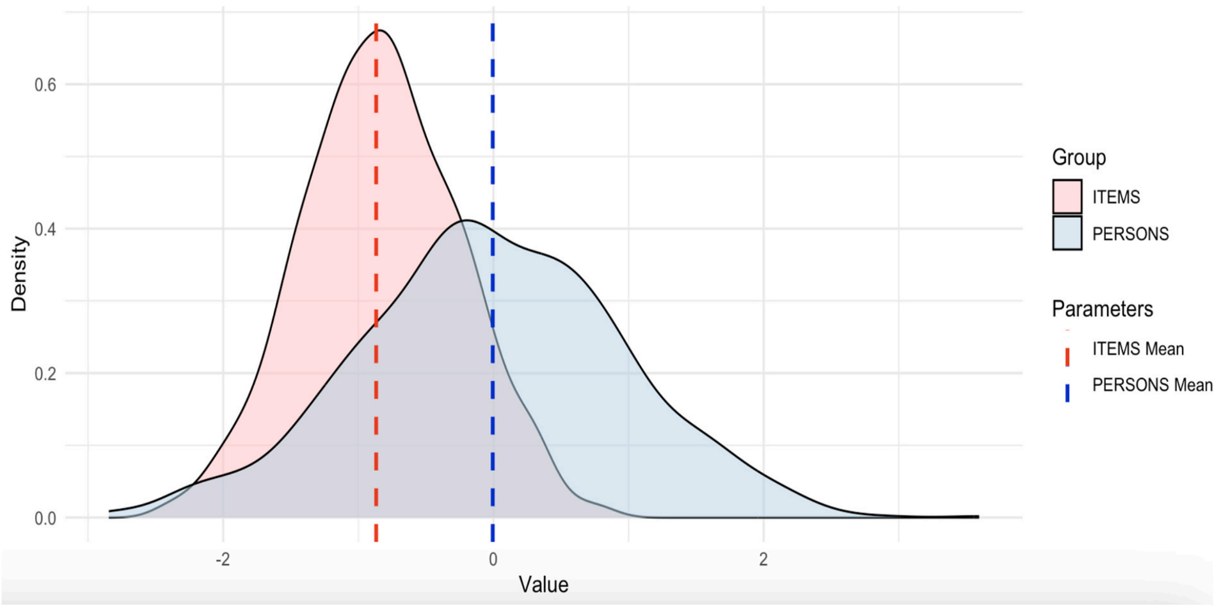

Fig. VII. Distribution of person and items

References

Adriaenssen, A., & Aertsen, I. (2015). Punitive Attitudes: Towards an Operationalization to Measure Individual Punitivity in a Multidimensional Way. *European Journal of Criminology, 12*(1), 92–112. https://doi.org/10.1177/1477370814535376

Aguilar-Jurado, J. A. (2018). Aproximación al Análisis de Las Actitudes Punitivas. *Revista Criminalidad, 60*(1), 95–110.

Aizpurúa, E. (2015). Delimitando el punitivismo. Las actitudes de los españoles hacia el castigo de los infractores juveniles y adultos. *Revista Española de Investigación Criminológica, 13*, 1–30.

Applegate, B. K., Cullen, F. T., & Fisher, B. S. (2002). Public Views toward Crime and Correctional Policies: Is There a Gender Gap? *Journal of Criminal Justice, 30*(2), 89–100. https://doi.org/10.1016/S0047-2352(01)00127-1

Armborst, A. (2017). How Fear of Crime Affects Punitive Attitudes. *European Journal on Criminal Policy and Research, 23*(3), 461–481. https://doi.org/10.1007/s10610-017-9342-5

Baker, F. B., & Kim, S.-H. (2017). *The Basics of Item Response Theory Using R*. Springer.

Baker, F. B. (2001). *The Basics of Item Response Theory*. ERIC.

Blalock, H. M. (1968). 'The Measurement Problem: A Gap between the Languages of Theory and Research'. *Methodology. Social Research*, 5–27.

Bollen, K., & Lennox, R. (1991). Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin, 110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Brand, P. A., & Anastasio, P. A. (2006). Violence-Related Attitudes and Beliefs: Scale Construction and Psychometrics. *Journal of Interpersonal Violence, 21*(7), 856–868. https://doi.org/10.1177/0886260506288934

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item Response Theory. *Annual Review of Statistics and Its Application, 3*, 297–321. https://doi.org/10.1146/annurev-statistics-041715-033702

Campregher, J., & Jeglic, E. L. (2016). Attitudes Toward Juvenile Sex Offender Legislation: The Influence of Case-Specific Information. *Journal of Child Sexual Abuse, 25*(4), 466–482. https://doi.org/10.1080/10538712.2016.1153558

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. *Clinical Therapeutics, 36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Carlsmith, K. M. (2008). On Justifying Punishment: The Discrepancy Between Words and Actions. *Social Justice Research, 21*(2), 119–137. https://doi.org/10.1007/s11211-008-0068-x

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*, 1–29.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *Journal of Statistical Software, 39*(8), 1.

Conrad, C. R., Croco, S. E., Gomez, B. T., & Moore, W. H. (2018). Threat Perception and American Support for Torture. *Political Behavior, 40*(4), 989–1009. https://doi.org/10.1007/s11109-017-9433-5

Conrad, K. J., Riley, B. B., Conrad, K. M., Chan, Y.-F., & Dennis, M. L. (2010). Validation of the Crime and Violence Scale (CVS) Against the Rasch Measurement Model Including Differences by Gender, Race, and Age. *Evaluation Review, 34*(2), 83–115. https://doi.org/10.1177/0193841X10362162

Costelloe, M. T., Arazan, C., & Stenger, M. (2018). Assessing the Effect of Social Science Education on Punitive Attitudes. *JSSE - Journal of Social Science Education, 17*(3), 88–99. https://doi.org/10.4119/jsse-883

Courtright, K. E., Mackey, D. A., & Packard, S. H. (2005). Empathy among College Students and Criminal Justice Majors: Identifying Predispositional Traits and the Role of Education. *Journal of Criminal Justice Education, 16*(1), 125–144. https://doi.org/10.1080/1051125042000333514

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Cullen, F. T., Clark, G. A., Cullen, J. B., & Mathers, R. A. (1985). Attribution, Salience, and Attitudes toward Criminal Sanctioning. *Criminal Justice and Behavior, 12*(3), 305–331.

Cullen, F. T., Cullen, J. B., & Wozniak, J. F. (1988). Is Rehabilitation Dead? The Myth of the Punitive Public. *Journal of Criminal Justice, 16*(4), 303–317. https://doi.org/10.1016/0047-2352(88)90018-9

Cullen, F. T., Jonson, C. L., & Nagin, D. S. (2011). Prisons Do Not Reduce Recidivism: The High Cost of Ignoring Science. *The Prison Journal, 91*(3_suppl), 48S–65S.

Davey, C. B., Mulrooney, K. J. D., & Watt, S. E. (2024). Exploring Individual-Level Predictors of Punitive Attitudes in Australia. *Psychology, Crime & Law*, 1–21.

Dillard, C. L., Salekin, R. T., Barker, E. D., & Grimes, R. D. (2013). Psychopathy in Adolescent Offenders: An Item Response Theory Study of the Antisocial Process Screening Device–Self Report and the Psychopathy Checklist: Youth Version. *Personality Disorders, Theory, Research, and Treatment, 4*(2), 101–120. https://doi.org/10.1037/a0028439

Dodd, S. (2018). The Punitive Woman? Gender Differences in Public Attitudes Toward Parole Among an Australian Sample. *International Journal of Offender Therapy and Comparative Criminology, 62*(10), 3006–3022. https://doi.org/10.1177/0306624X17739560

Durham, I. I. I., & Alexis, M. (1988). Crime Seriousness and Punitive Severity: An Assessment of Social Attitudes. *Justice Quarterly, 5*(1), 131–153. https://doi.org/10.1080/07418828800089651

Enns, P. K. (2014). The Public's Increasing Punitiveness and Its Influence on Mass Incarceration in the United States. *American Journal of Political Science, 58*(4), 857–872. https://doi.org/10.1111/ajps.12098

Ezquerra, P. (2025). Moral Networks: an analysis of punitive attitudes using Big Data. *Doctoral Dissertation (forthcoming)*. Cardiff University.

Falco, D. L., & Martin, J. S. (2012). Examining Punitiveness: Assessing Views Toward the Punishment of Offenders Among Criminology and Non-Criminology Students. *Journal of Criminal Justice Education, 23*(2), 205–232. https://doi.org/10.1080/10511253.2011.631931

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Person Statistics. *Educational and Psychological Measurement, 58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Gault, B. A., & Sabini, J. (2000). The Roles of Empathy, Anger, and Gender in Predicting Attitudes toward Punitive, Reparative, and Preventative Public Policies. *Cognition and Emotion, 14*(4), 495–520. https://doi.org/10.1080/026999300402772

Gerber, M. M., & Jackson, J. (2016). Authority and Punishment: On the Ideological Basis of Punitive Attitudes towards Criminals. *Psychiatry, Psychology and Law, 23*(1), 113–134. https://doi.org/10.1080/13218719.2015.1034060

Giacomantonio, M., & Pierro, A. (2014). Individual Differences Underlying Punishment Motivation. *Social Psychology, 45*(6), 449–457. https://doi.org/10.1027/1864-9335/a000211

Gibson, C. L., Ward, J. T., Wright, J. P., Beaver, K. M., & Delisi, M. (2010). Where Does Gender Fit in the Measurement of Self-Control? *Criminal Justice and Behavior, 37*(8), 883–903. https://doi.org/10.1177/0093854810369082

Giguère, G., & Bourassa, C. (2023). Do the Redundant and Locally Dependent Items of the LS/CMI Contribute in Any Meaningful Way to Its Reliability and Its Potential to Predict Criminal Recidivism? *International Journal of Offender Therapy and Comparative Criminology*. https://doi.org/10.1177/0306624X231212815, 0306624X231212815.

Giguère, G., Bourassa, C., & Brouillette-Alarie, S. (2023). Effect of the Differential Item Functioning (DIF) of LS/CMI Items with Convicted Men and Women. *Journal of Experimental Criminology, 20*(3), 761–785. https://doi.org/10.1007/s11292-023-09559-9

Giguère, G., Higgs, T., & Charette, Y. (2023). Gender Effects in Actuarial Risk Assessment: An Item Response Theory Psychometric Study of the LS/CMI. *Women & Criminal Justice, 0*(0), 1–18. https://doi.org/10.1080/08974454.2023.2186199

Goertz, G. (2006). *Social Science Concepts: A User's Guide*. Princeton University Press.

Gordon, R. A. (2015). Measuring Constructs in Family Science: How Can Item Response Theory Improve Precision and Validity? *Journal of Marriage and Family, 77*(1), 147–176. https://doi.org/10.1111/jomf.12157

Gottfredson, M. R., & Hirschi, T. (1990). *A General Theory of Crime*. Stanford University Press.

Graham, A., Cullen, F. T., & Link, B. G. (2025). *The Hidden Measurement Crisis in Criminology: Procedural Justice as a Case Study*. Cambridge University Press.

Grasmick, H. G., Tittle, C. R., Bursik Jr, R. J., & Arneklev, B. J. (1993). Testing the Core Empirical Implications of Gottfredson and Hirschi's General Theory of Crime. *Journal of Research in Crime and Delinquency, 30*(1), 5–29.

Guillermo, S., Zuniga, J., & Quiroz, A. D. (2021). The Role of Intergroup Threat in Support of Punitive Policies Toward Mexican Immigrants. *Hispanic Journal of Behavioral Sciences, 43*(3), 237–256. https://doi.org/10.1177/07399863211034669

Hambleton, R. K., Swaminathan, H., & Jane Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hickman, M. J., Piquero, N. L., & Piquero, A. R. (2004). The Validity of Niederhoffer's Cynicism Scale. *Journal of Criminal Justice, 32*(1), 1–13. https://doi.org/10.1016/j.jcrimjus.2003.10.001

Higgins, G. E. (2007). Examining the Original Grasmick Scale: A Rasch Model Approach. *Criminal Justice and Behavior, 34*(2), 157–178. https://doi.org/10.1177/0093854806290071

Hogan, M. J., Chiricos, T., & Gertz, M. (2005). Economic Insecurity, Blame, and Punitive Attitudes. *Justice Quarterly, 22*(3), 392–412. https://doi.org/10.1080/07418820500219144

Horstman, N. J., Bond, C. E. W., & Eriksson, L. (2021). Sentencing Domestic Violence Offenders: A Vignette Study of Public Perceptions. *Journal of Interpersonal Violence, 36*(21–22). NP11916–39.

Intravia, J. (2019). Investigating the Influence of Social Media Consumption on Punitive Attitudes Among a Sample of U.S. University Students. *International Journal of Offender Therapy and Comparative Criminology, 63*(2), 309–333. https://doi.org/10.1177/0306624X18786610

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement, 40*(8), 559–572.

Jennings, W., Farrall, S., Gray, E., & Hay, C. (2017). Penal Populism and the Public Thermostat: Crime, Public Punitiveness, and Public Policy. *Governance, 30*(3), 463–481. https://doi.org/10.1111/gove.12214

Johnson, C., & Raudenbush, S. W. (2006). A Repeated Measures, Multilevel Rasch Model with Application to Self-Reported Criminal Behavior. In *Methodological Issues in Aging Research*. Psychology Press.

Keller, L. B., Oswald, M. E., Stucki, I., & Gollwitzer, M. (2010). A Closer Look at an Eye for an Eye: Laypersons' Punishment Decisions Are Primarily Driven by Retributive Motives. *Social Justice Research, 23*(2), 99–116. https://doi.org/10.1007/s11211-010-0113-4

Kornhauser, R. (2015). Economic Individualism and Punitive Attitudes: A Cross-National Analysis. *Punishment & Society, 17*(1), 27–53.

Kruis, N. E., Ménard, K. S., Choi, J., Rowland, N. J., Frye, T., Kosaka, R., & Williams, A. (2023). Perceived Dangerousness Mediates Punitive Attitudes Toward Sex Offenders: Results From a Vignette Experiment. *Crime & Delinquency, 0(0)*. https://doi.org/10.1177/00111287231170106

Linacre, J. M. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

der Linden, V., & Wim, J. (2017). *Handbook of Item Response Theory: Volume 3: Applications*. CRC Press.

Liu, P., Nunn, R., & Shambaugh, J. (2018, December 19). The Economics of Bail and Pretrial Detention. *Brookings Institute*.

Llabrés, M., & Torres, C. R. (2020). Aportes al Debate Acerca de Vivir Sin Miedo. *Fronteras, 14*.

Loeffler, C. E., & Nagin, D. S. (2022). The Impact of Incarceration on Recidivism. *Annual Review of Criminology, 5*, 133–152.

Lord, F. M. (2012). *Applications of Item Response Theory to Practical Testing Problems*. Routledge.

Mackey, D. A., & Courtright, K. E. (2000). Assessing Punitiveness among College Students: A Comparison of Criminal Justice Majors with Other Majors. *The Justice Professional, 12*(4), 423–441. https://doi.org/10.1080/1478601X.2000.9959561

Maguire, E., & Johnson, D. (2015). The Structure of Public Opinion on Crime Policy: Evidence from Seven Caribbean Nations. *Punishment & Society, 17*(4), 502–530. https://doi.org/10.1177/1462474515604385

Mascini, P., & Houtman, D. (2006). Rehabilitation and Repression: Reassessing Their Ideological Embeddedness. *The British Journal of Criminology, 46*(5), 822–836. https://doi.org/10.1093/bjc/azl014

Mathesius, J., & Lussier, P. (2021). Structural Equation Modeling. In *The Encyclopedia of Research Methods in Criminology and Criminal Justice* (pp. 884–889). John Wiley & Sons, Ltd.

Mayhew, P., & van Kesteren, J. (2002). Cross-National Attitudes to Punishment. In *Changing Attitudes to Punishment*. Willan.

Metcalfe, C., & Cann, D. (2020). Arab Threat and Social Control: An Exploration of the Relationship Between Ethnic Attitudes and Punitiveness Among Israeli Jews. *International Journal of Offender Therapy and Comparative Criminology, 64*(5), 498–521. https://doi.org/10.1177/0306624X19895973

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An Introduction to Item Response Theory for Patient-Reported Outcome Measurement. *The Patient - Patient-Centered Outcomes Research, 7*(1), 23–35. https://doi.org/10.1007/s40271-013-0041-0

Nivette, A. E. (2016). Institutional Ineffectiveness, Illegitimacy, and Public Support for Vigilantism in Latin America. *Criminology, 54*(1), 142–175. https://doi.org/10.1111/1745-9125.12099

Opción Consultores. (2018). *Informe Opinión sobre Cadena Perpetua y Pena de Muerte para Delitos Graves - Marzo 2018*. Opción Consultores - Portal de Opinión Pública. Retrieved 15 January 2025 https://www.opcion.com.uy/opinion-publica/informe-opinion-sobre-cadena-perpetua-y-pena-de-muerte-para-delitos-graves-marzo-2018/.

Ortet-Fabregat, G., & Pérez, J. (1992). An Assessment of the Attitudes towards Crime among Professionals in the Criminal Justice System. *The British Journal of Criminology, 32*(2), 193–207.

Osgood, D. W., Finken, L. L., & McMorris, B. J. (2002). Analyzing Multiple-Item Measures of Crime and Deviance II: Tobit Regression Analysis of Transformed Scores. *Journal of Quantitative Criminology, 18*(4), 319–347. https://doi.org/10.1023/A:1021198509929

Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing Multiple-Item Measures of Crime and Deviance I: Item Response Theory Scaling. *Journal of Quantitative Criminology, 18*(3), 267–296. https://doi.org/10.1023/A:1016008004010

Ostaszewski, P., Uhl, A., Witkowska-Rozpara, K., & Woźniakowska, D. (2024). Punitiveness of Society and Criminal Policy in Six Central European Countries. *European Journal of Criminology, 21*(6), 929–956. https://doi.org/10.1177/14773708241260153

Payne, B. K., Gainey, R. R., Triplett, R. A., & Danner, M. J. E. (2004). What Drives Punitive Beliefs?: Demographic Characteristics and Justifications for Sentencing. *Journal of Criminal Justice, 32*(3), 195–206. https://doi.org/10.1016/j.jcrimjus.2004.02.007

Pechorro, P., DeLisi, M., Pacheco, C., Gonçalves, R. A., Maroco, J., & Quintas, J. (2023). Examination of Grasmick et al.'s Low Self-Control Scale and of a Short Version With Cross-Gender Measurement Invariance. *Crime & Delinquency, 69*(13–14), 2741–2764. https://doi.org/10.1177/00111287211073674

Piquero, A. R., MacIntosh, R., & Hickman, M. (2000). Does Self-control Affect Survey Response? Applying Exploratory, Confirmatory, and Item Response Theory Analysis to Grasmick et al.'s Self-control Scale. *Criminology, 38*(3), 897–930.

Piquero, A. R., MacIntosh, R., & Hickman, M. (2001). Applying Rasch Modeling to the Validity of a Control Balance Scale. *Journal of Criminal Justice, 29*(6), 493–505. https://doi.org/10.1016/S0047-2352(01)00112-X

Piquero, A. R., Macintosh, R., & Hickman, M. (2002). The Validity of a Self-Reported Delinquency Scale: Comparisons across Gender, Age, Race, and Place of Residence. *Sociological Methods & Research, 30*(4), 492–529.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/.

Ramirez, M. D. (2013). Punitive Sentiment. *Criminology, 51*(2), 329–364. https://doi.org/10.1111/1745-9125.12007

Ramirez, M. D. (2015). Racial Discrimination, Fear of Crime, and Variability in Blacks' Preferences for Punitive and Preventative Anti-Crime Policies. *Political Behavior, 37*(2), 419–439. https://doi.org/10.1007/s11109-014-9285-1

Reise, S. P., & Waller, N. G. (2009). Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology, 5*(1), 27–48.

Rocque, M., Posick, C., & Zimmerman, G. M. (2013). Measuring Up: Assessing the Measurement Properties of Two Self-Control Scales. *Deviant Behavior, 34*(7), 534–556. https://doi.org/10.1080/01639625.2012.748619

Rozin, P. (2001). Social Psychology and Science: Some Lessons From Solomon Asch. *Personality and Social Psychology Review, 5*(1), 2–14. https://doi.org/10.1207/S15327957PSPR0501_1

Samejima, F. (1968). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *ETS Research Bulletin Series, 1968*(1), i–169.

Sanjurjo, D., Trajtenberg, N., & del Castillo, F. (2021). Policing in Uruguay. History, Modernization, and Features. In J. Mbuba (Ed.), *Global Perspectives in Policing and Law Enforcement*. Rowman & Littlefield.

de Sierra, L. P. (2019). Desafíos ante el avance del gobierno punitivo en Uruguay. *Escenarios, 30*.

Silver, J. R., & Silver, E. (2017). Why Are Conservatives More Punitive than Liberals? A Moral Foundations Approach. *Law and Human Behavior, 41*(3), 258–272. https://doi.org/10.1037/lhb0000232

Simmler, M., Stempkowski, M., & Markwalder, N. (2021). Punitive Attitudes and Victimization among Police Officers in Switzerland: An Empirical Exploration. *Police*

Practice and Research, 22(2), 1191–1208. https://doi.org/10.1080/15614263.2019.1697264

Singer, A. J., Chouhy, C., Lehmann, P. S., Stevens, J. N., & Gertz, M. (2020). Economic Anxieties, Fear of Crime, and Punitive Attitudes in Latin America. *Punishment & Society, 22*(2), 181–206. https://doi.org/10.1177/1462474519873659

Slyke, V., Shanna, R., Benson, M. L., & Virkler, W. M. (2018). Confidence in the Police, Due Process, and Perp Walks. *Criminology & Public Policy, 17*(3), 605–634. https://doi.org/10.1111/1745-9133.12380

Socia, K. M., Rydberg, J., & Dum, C. P. (2019). Punitive Attitudes Toward Individuals Convicted of Sex Offenses: A Vignette Study. *Justice Quarterly, 0*(0), 1–28. https://doi.org/10.1080/07418825.2019.1683218

Spiranovic, C. A., Roberts, L. D., & Indermaur, D. (2012). What Predicts Punitiveness? An Examination of Predictors of Punitive Attitudes towards Offenders in Australia. *Psychiatry, Psychology and Law, 19*(2), 249–261. https://doi.org/10.1080/13218719.2011.561766

Stalans, L. J. (2013). Measuring Attitudes to Sentencing. In *Changing attitudes to punishment* (pp. 33–50). Willan.

Subrayado. (2019). *El 54% está a favor de la reforma constitucional que impulsa Larrañaga. March 27.* subrayado.com.uy.

Sweeten, G. (2012). Scaling Criminal Offending. *Journal of Quantitative Criminology, 28*(3), 533–557. https://doi.org/10.1007/s10940-011-9160-8

Tam, K.-P., Al, A., & Leung, A. K.-Y. (2008). Attributionally More Complex People Show Less Punitiveness and Racism. *Journal of Research in Personality, 42*(4), 1074–1081. https://doi.org/10.1016/j.jrp.2007.11.002

Tittle, C. R. (2018). *Control Balance: Toward a General Theory of Deviance*. Routledge.

Torres Irribarra, D., & Freund, R. (2014). Wright Map: IRT Item-Person Map with ConQuest Integration. *R Package. Retrieved January, 10*, 2015.

Trajtenberg, Nicolás, Pablo Ezquerra, and Matthew Williams. 2024. '"Lock Them up and Throw Away the Key": An Evaluation of the Structure of Punitive Attitudes'. Psychiatry, Psychology and Law 0(0):1–27. doi: https://doi.org/10.1080/13218719.2023.2296476.

Viney, W., Waldman, D. A., & Barchilon, J. (1982). Attitudes Toward Punishment in Relation to Beliefs in Free Will and Determinism. *Human Relations, 35*(11), 939–949. https://doi.org/10.1177/001872678203501101

Ward, J. T., Gibson, C. L., Boman, J., & Leite, W. L. (2010). Assessing the Validity of the Retrospective Behavioral Self-Control Scale: Is the General Theory of Crime Stronger Than the Evidence Suggests? *Criminal Justice and Behavior, 37*(3), 336–357. https://doi.org/10.1177/0093854809359673

Wilson, M. (2023). *Constructing Measures: An Item Response Modeling Approach.* Routledge.

Wind, S. A. (2016). Examining the Psychometric Quality of Multiple-Choice Assessment Items Using Mokken Scale Analysis. *Journal of Applied Measurement, 17*(2), 142–165.

Zanon, C., Hutz, C. S., Yoo, H. H., & Hambleton, R. K. (2016). An Application of Item Response Theory to Psychological Test Development. *Psicologia: Reflexão e Crítica, 29*. https://doi.org/10.1186/s41155-016-0040-x