ARTICLE IN PRESS

+ MODEL

Available online at www.sciencedirect.com

ScienceDirect

International Journal of Law, Crime and Justice xx (2015) 1–20

International Journal of Law, Crime and Justice

www.elsevier.com/locate/ijlcj

Refining the measurement of consistency in sentencing: A methodological review

Jose Pina-Sánchez [a,*], Robin Linacre [b,1]

[a] *London School of Economics, Department of Statistics, Columbia House, Houghton Street, London WC2A 2AE, UK*
[b] *The Sentencing Council for England and Wales, The Royal Courts of Justice, East Block, Room EB16, The Strand, London WC2A 2LL, UK*

Abstract

The importance of improving consistency in sentencing has been underscored by institutional reforms in a number of jurisdictions. However, the effectiveness of these policy changes has not been clearly measured. To a certain extent this is due to the methodological confusion reflected by the multiplicity of methods that have been used in the study of consistency in sentencing. Here we review and categorise all of the quantitative methods that have been used to measure consistency in the literature. Our classification differentiates methods based on characteristics such as their robustness, the type of data they require, or whether they are amenable to comparisons in time or across jurisdictions. In this way the paper has a twofold contribution: it simplifies the implementation of future empirical analyses on consistency and facilitates their critical interpretation.

Keywords: Sentencing; Consistency; Disparity; Measurement; Quantitative

* Corresponding author. Tel.: +44 (0)2079556717.
E-mail addresses: j.pina-sanchez@lse.ac.uk (J. Pina-Sánchez), robin.linacre@sentencingcouncil.gsi.gov.uk (R. Linacre).
[1]Tel.: +44 020 7071 5793.

http://dx.doi.org/10.1016/j.ijlcj.2015.06.001
1756-0616/

1. Introduction

The assurance that like cases will be treated alike regardless of where, when, or who is sentencing them is a fundamental principle of justice. Furthermore, in addition to being a goal in its own right, consistency in sentencing is also associated with other desirable effects. For example, it fosters public confidence in sentencing, helps to establish a common understanding of the consequences of crime, and promotes the legitimacy of the criminal justice system. However, the means by which consistency in sentencing can be achieved are neither clear-cut nor uncontroversial.

Over the years, critics of the sentencing process in western jurisdictions have contended that unrestrained discretion in the hands of sentencers leads to inconsistency in sentencing.[2] This has led many jurisdictions to introduce greater structure for sentencers, usually in the form of sentencing guidelines. The exact nature of guidelines varies widely; in simple terms we could position them along a continuous scale reflecting the extent to which judicial discretion is restricted (Reitz, 2013). For example, across the US many states and the federal system employ relatively rigid guidelines, often in the form of a sentencing grid (see Frase, 2005a), where types of offences are associated with specific sentence outcomes. In contrast Scandinavian countries have typically issued 'guidance by words' (see Wandall, 2006), while in England and Wales the Sentencing Council has devised a system of guidelines which lies between these two paradigms (see Dhami, 2013; Roberts, 2013a).

Despite the considerable attention attracted by these reforms, the effectiveness of different approaches to structuring discretion remains an open question. American scholars have led the research efforts, although significant work did not begin until the late 90s, and struggled to reach definitive conclusions. This view was first expressed by Tonry (1996) and Austin et al. (1996), as recounted by Hofer et al. (1999): "*Evaluations of state systems have been few, and independent evaluations almost non-existent*"; and "*The past 20 years have produced many accusations but few studies documenting the misuse of discretion by judges*" (Hofer et al., 1999, p. 262).

Along the new century we have seen a substantial increase in the number of contributions, many of them investigating the consequences of the 'Booker/Fanfan' reform of the federal guidelines. However, in spite of these new contributions – addressing a specific case study – no definitive answers have been found (Engen, 2011; Sebba, 2013)[3]:"*It is difficult to comment on the impact of sentencing guidelines on sentencing disparity because there simply is little empirically rigorous research examining the effects of actual policy changes*" (Engen, 2011, p. 1139). Furthermore, this is not a peculiarity of the US. Recent reviews looking at consistency in other jurisdictions, such as Pina-Sánchez and Linacre (2013) in England and Wales and Krasnostein and Freiberg (2013) in Australia, have noted similarly inconclusive results.

We believe that there are two fundamental reasons why so little is known about such a highly debated topic. First and foremost, there is not enough good data available (Krasnostein and Freiberg, 2013; Schanzenbach and Tiller, 2008).[4] The difficulty of convincing the judiciary

[2] Frankel's (1972) influential essay could be considered the spark that ignited the debate on sentencing reform when he claimed that unstructured discretion leads to 'lawlessness' in sentencing.

[3] "*scholars perceive the research findings to be equivocal as to the extent to which the reforms have achieved their objectives*" (Sebba, 2013, p. 257).

[4] "*The underlying antipathy to social science data in the courts has limited their utility in identifying patterns of sentencing in commonly occurring crimes*" (Krasnostein and Freiberg, 2013, p. 277).

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

of the benefits of recording and disseminating detailed data at the individual sentence level is a major obstacle – the usual practice is to publish aggregated data on the different cases being sentenced. Even where record level data is available, such as that published by the Sentencing Council for England and Wales, important details of the case (e.g. judge/court identifiers) tend to be omitted. Very little can be done to resolve this problem besides expressing our disappointment for the perpetuation of such practices despite the substantial progress in open data in other domains.

A second problem that has hindered the progress of research on the topic is the lack of agreement on what is meant by 'consistency in sentencing' and the way it should be measured (Tonry, 1996; Casey and Wilson, 1998; Hofer et al., 1999),[5] "*Divergent points of view are common in the arena of sentencing policy. But such a range of opinion about an important empirical matter indicates a failure of research to provide objective, quantified answers to these essentially factual questions*" (Hofer et al., 1999, p. 263). The lack of clarity of the concept is reflected in the multiplicity of terms that have been used to refer to consistency or lack of thereof (Sebba, 2013)[6]; e.g. 'equality', 'discrepancy', 'uniformity', 'disparity' are variously applied. This is recognised by the statutory authorities responsible for promoting consistency in sentencing in the US and in England and Wales, which have confirmed that there is not a unique definition of what should be understood by consistency in sentencing[7]; "*While there is widespread agreement that unwarranted disparity should be eliminated, there is less agreement on how to define it*" (United States Sentencing Commission, 2004, p. 79).

The ambiguity surrounding the concept of consistency together with the reliance on suboptimal data complicates the operationalisation of the concept. As a result, very different methods for measuring consistency have been proposed in the literature. These methodologies are highly variable in terms of scope and validity. Regrettably, the nuances and applicability of the various measures are seldom made clear, rendering the study of consistency highly idiosyncratic, and any meta-analysis or systematic attempt to compare levels of consistency across time or jurisdictions unreliable.[8]

The aim of this article is to present a clearer research framework to facilitate and promote the better measurement of consistency in sentencing. We intend to do so by undertaking a theoretical and a methodological review of the concept. In the following section we lay the foundations of what is understood by consistency, stressing the differences with often intertwined concepts such as proportionality, discrimination and uniformity. We do not seek to end the philosophical debate on the precise meaning of these concepts – instead, we wish to

[5] "*such research has been rife with methodological limitations not least of which is the failure to quantify or appropriately define disparity. This calls into question the true level of disparity within the system.*" (Casey and Wilson, 1998, p. 237). "*Although a number of studies have been conducted that appear to demonstrate the existence of various forms of disparity, there is a dearth of conclusive empirical evidence of the nature and extent of unjustified disparity in Australia. This is due partly to the difficulty of conceptualising and operationalising the notion of 'unjustified disparity*" (Krasnostein and Freiberg, 2013, p. 272).

[6] "*Some of these constraints and confusions are embedded in the ambiguities of the concepts incorporated in the sentencing discourse, including commonly used terms such as 'disparities' (and 'legal' and 'extralegal' variables)*" (Sebba, 2013, p. 240).

[7] "*there is no universally accepted definition of consistency in sentencing*" (Sentencing Council for England and Wales 2011, p. 1).

[8] "*The first step toward reaching reasonable conclusions about the success of the guidelines is to understand the different methods and the questions that can be answered by each*" (Hofer et al., 1999, p. 264).

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

provide definitions that enable the measurement of consistency that approximate as closely as possible the qualitative meanings of these concepts.

After this conceptual analysis, we describe the different elements of consistency that can be considered when operationalising the concept. Finally, we review and categorise the methods that have been used in the literature to measure consistency in sentencing. This classification distinguishes methods based on characteristics such as their robustness, the type of data they require, and whether they permit comparisons in time or across jurisdictions.

In spite of the clear benefits, there have been few methodological reviews in the area, the exceptions being Anderson et al. (1999), Hofer et al. (1999) and Pina-Sánchez and Linacre (2013). These three papers include a comparison of methods preceding the development of their analyses, thus limiting the attention they could give to such comparisons, which are closer to literature reviews than to systematic comparisons. In addition, Hofer et al. (1999), the most detailed of these reviews, is now over a decade old. We extend these authors' work by reporting a more exhaustive review of methods and by updating it with new ones that have been used in recent years.

2. Defining consistency

To be able to measure the degree of consistency in sentencing we first need a clear idea of what consistency means. The standard phraseology 'the extent to which like cases are treated alike' captures the essence of the concept, but it also disguises its complexities. Here, we approach this problem by examining each of the parts of the proposition separately. Let us consider first what is meant by 'like cases'.

Like cases are those that share the same legal factors, which we define as the case characteristics that may legitimately affect the sentencing decision. These will vary across jurisdictions and offences, but will typically include aspects of harm and culpability, personal mitigation, and procedural factors such as guilty plea discounts. From this premise it follows that cases should not be differentiated on the basis of non-legal factors (i.e. those that should not have an effect on the sentence outcome). Again, what is to be considered as a non-legal factor may vary across jurisdictions and offences,[9] however, the identity of the judge, the geographical area where the sentence was passed, or the offender's race, gender, religion, sexual orientation, or social class, represent some of the factors that could be generally understood as non-legal.

Regarding the second part of the proposition, '… are treated alike', two interpretations are possible. Some understand that the *approach* to sentencing (i.e. the process followed to deduce the sentence outcome) should be consistent across sentencing decisions. This is the definition used by the Sentencing Council for England and Wales[10]. Others understand that it is the sentence *outcomes* amongst like cases that must be alike. These two views have been used in the literature to differentiate between consistency of approach and consistency of outcome (Hola, 2012; Krasnostein and Freiberg, 2013).

[9]For example, age or lack of maturity is considered a mitigating factor in the 2011 England and Wales Assault Guideline, while the Crown Prosecution Service does not consider any differential treatment on those basis. Similarly, the code of law in Spain defines cases as domestic violence only when perpetrated by a man, whereas in England and Wales offenders of domestic violence are treated equally regardless of their gender.

[10]Sentencing Council press release 01 December 2010 http://sentencingcouncil.judiciary.gov.uk/media/reminder.htm.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

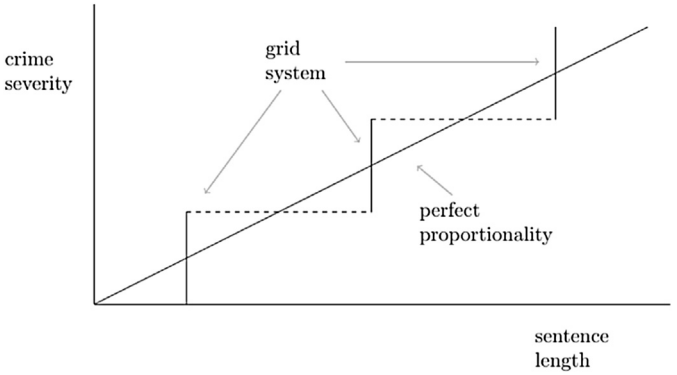

Fig. 1. Uniformity as the result of fixing sentence outcomes.

Our opinion is that these different views should not produce competing definitions of consistency since a similar approach to sentencing should result in alike cases receiving similar sentence outcomes. However, the requirement is only that similar outcomes should prevail where *all* the legal factors are similar. When some of those legal factors are overlooked consistency of outcome is normally used as a synonym to the concept of uniformity, which refers to the use of similar sentences outcomes for cases that should be treated differently. The distinction between consistency and uniformity can be subtle but it must be underlined as it has important policy and methodological implications As argued by Alschuler (2005) in his critique of the US Federal Guidelines experience, sentencing guidelines that are too restrictive (i.e. those that fail to account for all of the legal factors that legitimately define a case) might simply promote uniformity and in so doing hinder the principle of proportionality.

This is shown in Fig. 1, where we compare a perfectly proportional sentencing system – understood as that where the severity of sentences reflects all of the legal factors present in each case and the nuances that could be possibly considered – with one structured by a strict grid system, where cases are exclusively defined by a non-comprehensive list of prescribed legal factors.[11] We can see how under such a grid system most cases will receive either harsher or more lenient sentences than what they ought to have as a result of sentences being forcedly grouped within a limited number of outcomes.[12]

We continue our conceptualisation of inconsistency by discussing its relationship with discrimination. These two concepts are often conflated, which contributes to the overall confusion as to what should be understood by consistency. Discrimination refers to differences in sentences associated with one or more characteristics of the defendant,[13] e.g. ethnicity, gender, social class, etc. It is unclear where this list ends, arguably any personal characteristics

[11]This is just a hypothetical case since even the most rigid grid-based systems allow for departures from their prescribed outcomes (see for example Section D from the Minnesota Sentencing Guidelines Commission, 2014).

[12]To simplify the matter we have assumed that the severity of the sentence (measured by the sentence length) is a product of the seriousness of the crime. That is, the graph depicts a sentencing system entirely based on a retributive principle. However, the point remains under a more realistic scheme where other principals such as rehabilitation and the characteristics of the offender are taken into account; imposing a limited number of sentence outcomes hinders proportionality.

[13]Stolzenberg and D'Alessio (1994) and Bushway and Piehl (2001) use the term unwarranted disparity to refer to what we have defined here as discrimination.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

that can be associated with a sentence outcome after controlling for the legal factors defining the offence should be understood as discrimination.

Discrimination constitutes a violation of the principle of consistency because it entails treating 'like' cases differently. As such, the concept of discrimination is subsumed within the broader one of inconsistency. This view reflects Frankel's (1972) characterisation of discrimination as an element of 'lawlessness in sentencing', but contrasts with other views expressed in the literature where the two concepts are considered separately (Spohn, 2000, 2002; Ostrom et al., 2008). The latter are challenged in Fig. 2, which we will also use to establish how the concept of consistency should be considered to be independent from that of proportionality.

Proportionality concerns whether the severity of the sentence corresponds to the seriousness of the offence, and whether the relative severity of sentencing for different offence types is equitable. As such, a measure of proportionality may seek to reflect how far sentencing decisions deviate from the 'most appropriate' sentence for any offence. A measure of consistency, on the other hand, should only reflect how far sentencing decisions vary amongst similar cases. Hence, consistency is a positive concept whereas proportionality is normative; or, in statistical terms, lack of proportionality reflects a problem of bias while consistency relates to precision.

In consequence, studies seeking to measure the degree of consistency should not be involved in answering questions such as what is the purpose of sentencing (utilitarian, retributive, etc.) or what is an appropriate sentence. However, we have detected a number of studies (Britt, 2009; Bushway and Piehl, 2001; Engen, 2011; Fischman and Schanzenbach, 2012; Hofer, 2007) that make these types of conceptual mix-ups. Take for example the following sentence from Engen (2011), which reflects perfectly our point on how the confusion surrounding the concept of consistency has contributed to the scarcity of empirical evidence on the topic: "*how are we to evaluate the exercise of judicial discretion, or changes in sentencing disparity, relative to guidelines that many observers, including federal judges, believe are unjust?*" (Engen, 2011, p. 1145).

The differences amongst the concepts of proportionality, discrimination and consistency described here can be seen more clearly using a diagram. In Fig. 2 we have represented ten

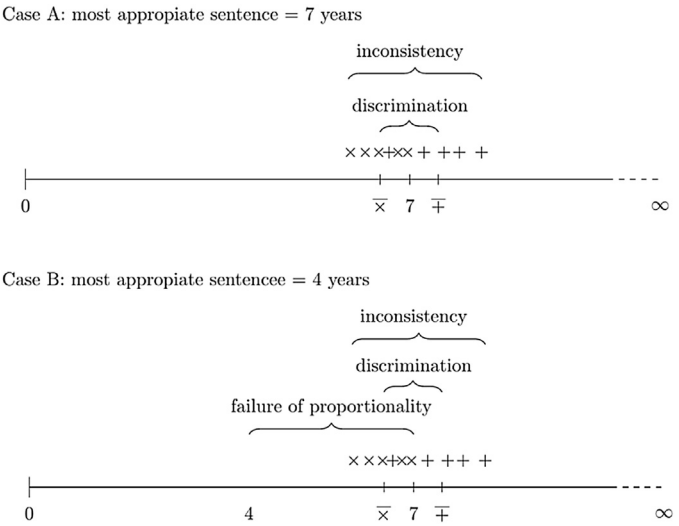

Fig. 2. Proportionality, consistency and discrimination.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

hypothetical sentences given for two specific cases. In the first case the 'most appropriate' sentence should have been seven years of custody, while in the second it should have been four years. The '+' represent sentences passed on offenders from a minority group, the '×' represent cases with offenders belonging to the dominant ethnic group, $\overline{+}$ and $\overline{\times}$ are their respective means.

Under this setting we can measure failures of proportionality as the difference between the most appropriate sentence and the average sentence outcome, which in Case A would be null but in Case B it would be equal to three. Discrimination can be defined as the difference between $\overline{+}$ and $\overline{\times}$, while inconsistency is represented by the dispersion of the sentence outcomes. Notice how discrimination is taken as an element of inconsistency, and how the measurement of consistency (or lack of thereof) does not rely on knowing what is the most appropriate sentence.

3. Operationalising consistency

The arguments laid out in the previous section contribute to establishing a clearer theoretical framework on what consistency is and is not. In this section we will use many of the features of consistency discussed previously to operationalise the concept. Specifically, we consider the underpinnings that research designs used to measure consistency in sentencing should contemplate.

We start by discarding methods that seek to measure consistency in other legal processes than sentencing, i.e. studies assessing consistency in the decisions of defense attorneys, prosecutors, probation and law enforcement officers (Alschuler, 2005), or at the stages of plea-bargaining and fact-finding (Starr and Rehavi, 2013). All of these tend to be discretionary processes which contribute to the overall level of inconsistency of the legal system. However, for reasons of space, their analysis is beyond the scope of this paper. In addition, we restrict our review to quantitative methods measuring outcome consistency. Qualitative methods (such as focus groups or interviews) are extremely valuable to ascertain the underlying processes in judges' deliberations and gain knowledge on procedural consistency[14] (Hola, 2012). However, they suffer from serious design flaws in the form of generalisability and replicability. The qualitative nature of the results makes it very difficult to obtain reliable measurements that could allow making comparisons with other countries or periods of time.

For the same reasons, we discarded quantitative methods used to provide evidence of inconsistency without measuring it. Some examples are studies that have used regression analysis to detect whether some legitimate legal factors are not having the expected effect (Kramer and Ulmer, 2002; Britt, 2009), or whether discrimination arises in the form of irrelevant personal characteristics having an effect in determining sentence outcomes (Albonetti, 1997, 2002; Everett and Wojtkiewicz, 2002; Mustard, 2001; Pasko, 2002; Stacey and Spohn, 2006; Steffensmeier and Demuth, 2000). Several of such studies indicated an interest on measuring disparities when they are specifically looking at different types of discrimination. By discarding these studies the number to be reviewed is reduced substantially,[15] since much greater academic efforts have been dedicated to the study of discrimination than to that of consistency in general.

[14] For example Davies et al. (2002) and Hough et al. (2003), ran focus groups with Crown Court judges and magistrates in England and Wales to study their views on how different offences should be sentenced. Both studies found significant divergences in the factors taken into account when imposing custodial sentences.

[15] "*Hundreds of such studies have been conducted during the past three decades (e.g. Hagan, 1974; Spohn, 2000), and they represent by a long mile the 'modal' approach to studying race and sentencing*" (Baumer, 2013, p. 234).

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

As a result of these restrictions we have narrowed down the list of methods to be reviewed to just those that can be used to measure outcome consistency in sentencing. Most of these methods share a common approach – although rarely expressed in the terms that we use here. They study the variability of sentence outcomes ($Var(O)$), which is considered to be formed by a legitimate ($Var(L)$) and illegitimate ($Var(I)$) component,

$$Var(O) = Var(L) + Var(I) \quad (1)$$

Specifically, $Var(L)$ reflects the variability across sentences that we would expect to observe for different cases based on the principle of proportionality, whereas $Var(I)$ reflects unwarranted disparities, i.e. variability across sentences that cannot be attributed to the legal factors defining the case.[16] Under this setting we can determine the overall level of consistency as the ratio $Var(I)/Var(L)$. Furthermore, returning to the concept of uniformity we can observe how a reduction of the possible sentence outcomes (i.e. a reduction of $Var(O)$) will not ensure an improvement of consistency as it could take the form of reduced $Var(L)$.

To generate measurements of consistency most methods seek to isolate $Var(I)$ in Eq. (1). This is achieved either using random allocation of cases across judges or by controlling for all of the relevant legal factors that define a case. The former approach relies on quasi-experimental data, which as we will see in the next section requires simpler analyses and fewer assumptions. However, the practice of random allocation of legal cases to judges is not widespread, making most sentencing data available not experimental but observational.

On the other hand, the reliance on statistical controls to isolate $Var(I)$ is a controversial issue. In particular, it is not clear what factors should be considered in the definition of a case, which gives rise to a normative debate that is difficult to resolve. The lack of consensus in this regard was summarised by Cole (1997) under the expression 'the empty idea of sentencing disparity', by which the author meant that the concept of consistency is only meaningful once it has been agreed on the case characteristics that define cases as alike or different.[17]

This pessimistic view can be challenged in certain circumstances, for example, in studies of offences regulated by grid-based guidelines, where the list of legal factors defining a case is finite and pre-established. To a certain extent, the same could be argued about semi-restrictive systems such as the England and Wales guidelines, which offer detailed – although not comprehensive – lists of applicable legal factors (see Sentencing Council, 2014). The problem arises when dealing with common law jurisdictions that are not structured by any kind of guidelines, such as the Australian state/territory and Federal jurisdictions. In these cases the normative impasse regarding the choice of legal factors could be overcome by reviewing relevant caselaw and using those legal factors noted by previous sentencers.

Finally, both experimental and observational data can be enriched when a hierarchical structure is also recorded. In particular, there are three levels that are relevant for the study of consistency: sentences (at level 1) can be grouped together according to the sentencing judge

[16] "*Unwarranted disparity is eliminated when sentencing decisions are based only on offense and offender characteristics related to the seriousness of the offense, the offender's risk of recidivism, or some other legitimate purpose of sentencing*" (USSC, 2004, pp. 80).

[17] "*[…] no agreement exists regarding the overarching normative theory of the guidelines or the proper priorities among the various purposes of sentencing. Different commentators and judges are free to choose their own and cry Disparity! whenever sentences do not conform to their favoured view*" (Hofer, 2007, p. 452).

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

(level 2), and in turn judges can be grouped within court centres (level 3). The presence of such hierarchical structure will determine the capacity of the method to assess a final distinction between forms of consistency that we will establish: *inter* and *intra-judge* consistency. The former refers to judges sentencing differently from one another, whereas the latter reflects inconsistencies in sentences passed by the same judge (Brantingham, 1985).[18]

Traditionally, most studies on consistency have solely focused on the inter-judge dimension. For example, Anderson et al. (1999) define consistency as "*[…] the variation in sentence that would result if a single offender were processed through the criminal justice by every possible combination of sentencing-decision makers*" (p. 3). However, this definition ignores the existence of intra-judge inconsistencies, which could arise as a result of different factors affecting the performance of judges – including trivial factors such as whether the judge has just had lunch (Danziger et al., 2011), or fluctuations of their workload across longer timespans.

As we will see in the following section some research designs offer the possibility of looking at either inter or intra-judge disparities separately. However, the ideal solution would be to use methods that capture inconsistencies from both dimensions simultaneously. This is what we refer to as *system* inconsistency. Lastly, when no judge level identifier is present in the dataset, some methods can also rely on the two level structure of sentences within courts.[19] This has the deficiency of producing less accurate measures of consistency. The higher the number of judges operating in the same court the coarser the measure of consistency.[20]

4. Review of methods

In this Section we proceed to review eleven methods used in the literature that comply with the criteria laid out before (measuring outcome consistency at sentencing). We will classify these methods based on the dimension of consistency they can address (inter, intra, or system consistency) and the various data types just discussed. Furthermore, to facilitate reliable comparisons between studies we will also identify the degrees of robustness, replicability and generalisability that should be attributed to different research designs.

By robustness we refer to the validity of the assumptions underpinning the method, and its resilience to their violation. Replicability refers to the facility with which the study can be rerun across time and/or jurisdictions, and the extent to which findings from replicated studies are comparable. Generalisability refers to the extent to which results are representative of the practice of sentencing in the entire jurisdiction. A vast majority of studies fail on this criterion as they tend to limit their analysis to sentence outcomes in the form of either custodial rate (a binary process indicating whether the offender is sent to custody) or sentence length (censoring out non-custodial offences).[21] Regrettably, both of these measures will exclude large arrays of possible distinctions in sentence outcomes amongst offenders.[22] Additionally, because of how they are

[18] Brantingham (1985) refers to this as first and second-order disparities, respectively.

[19] For example, that is the structure present in data from the Crown Court Sentencing Survey.

[20] On the other hand, some sources of inconsistency are better detected from the court level, e.g. Fearn (2005) found that districts with higher unemployment rate in the US pass harsher sentences.

[21] Ostrom et al. (2008) refer to the study of consistency that focuses on custodial rate or sentence length as consistency in location or duration, respectively.

[22] The statistical analysis of sentencing practices has benefited enormously in the last two decades from the incorporation of methods such as the Tobit (Albonetti, 1997, 1998) and two-stage Heckman models (Peterson and Hagan, 1984), which can be used to specify both the event of custodial sentence and sentence length simultaneously (for a good review see Bushway and Piehl, 2001).

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

designed, many methods can exclusively refer to particular courts or types of offences. This lack of generalisability can be quite problematic since the level of consistency may vary amongst them. For example, we may expect that sentences for common types of offences, for which there exist large bodies of caselaw and often sentencing guidelines, will be more consistent than sentences for unusual offences.

Table 1 sums up the characteristics (in columns) defining each of the eleven methods (in rows) that we have selected. The first two characteristics indicate the source of inconsistency that the method can assess and the type of data required, while for the other four characteristics we use bullet points where we believe that a method has a comparative advantage. In what follows these differences between methods will be spelt out in more detail.

4.1. Experimental simulations

This research design involves distributing hypothetical case-files to different judges who are then requested to provide an appropriate sentence for each case. Through these *simulations* we can directly look at inter-judge consistency by taking a measure of dispersion (such as a standard deviation)[23,24,25] of all the sentences passed by the participant judges on the same case.

Some examples of research using *simulations* are: the Second Circuit sentencing study (1981) by the US Department of Justice, the studies carried out for the Supreme Court of the United Kingdom for their own evaluations, or Davies and Tyrer (2003), where 51 judges from twelve courts in England and Wales participated in 'sentencing' five domestic burglary scenarios.

The main drawback of this method lies in its lack of generalisability since the measures of consistency apply to very specific cases.[26] Inferring an overall measure of consistency from this type of evidence is highly tentative, especially if we take in consideration that this method can only detect inter-judge − not system − consistency.

The general validity of the method is also a major issue since it uses simulated data. As Anderson et al. (1999) indicate, it is very difficult for a simulation to reconstruct in full depth the information available to a judge in a real case. Furthermore, it is not clear that judges dedicate the same attention to simulated cases than they do to real ones. For example, it could be argued that judges will dedicate more attention than they would normally do if they feel that the experiment is used to assess their performance. In addition, it is important to notice that in *simulations* the legal factors to consider are already identified and the only disparities that can be captured are those that arise from different treatments of like cases. This is not the case for the other methods that we review here, which are based on real sentences and as a result can also capture disparities due to the consideration of non-legal factors such as race, for example.

[23]More formally, taking l to represent the sentence lengths passed by J number of participant judges represented by $j = 1, 2, \ldots, J$, a measure of (in)consistency can be calculated as follows, $S = \sqrt{1/J - 1\sum_{j=1}^{J}(l_j - \bar{l})^2}$.

[24]Here and in the rest of the methods reviewed we assume that the interest is in consistency on duration (analysis of sentence length) for reasons of space and because of the simpler formalisation of methods using continuous instead of categorical data.

[25]Normally the natural logarithm of l is taken to run analyses in sentence length both to normalise the variable which otherwise would be right-skewed and to better reflect the idea that the effect in severity of increasing sentence length by one year is higher in the first quantiles of the distribution than in the last ones. That is, the severity of increasing the sentence length by one year is higher when the base sentence is six months than when it is twenty years.

[26]This situation can be partially improved taking the average of the standard deviations obtained for every case studied and using weights to reflect how frequent each case is in the jurisdiction of study. For a k number of cases-files denoted by $k = 1, 2, \ldots, K$, this more representative measure of consistency can be expressed as, $\widehat{C} = \sum_{k=1}^{K}(n_k/NS_k)$, where the weights are defined as the sampling fraction of each case; that is, the ratio of the frequency of one particular case in the jurisdiction of study, n_k, over the sum of those frequencies for all the cases included in the simulation, N. However, this solution is only partial as there is a limit in the number of cases that sentencers can consider before a point where research fatigue is reached, after which the validity of reports can be compromised.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

Table 1
Comparison of research designs.

Method	Source of (in) consistency	Data requirement	Validity/ robustness	Generalisability	Across time comparisons	Across jurisdictions comparisons
4.1. Experimental simulations	Inter	Experimental			•	•
4.2. Randomised cases	Inter & intra	Experimental	•			
4.3a. Conditional comparisons	System	Observational		•	•	•
4.3b. Exact matching	System	Observational		•		•
4.4a. Cross-sectional analysis of residuals	System	Observational		•		
4.4b. Longitudinal analysis of residuals	System	Observational		•	•	
4.5a. Aggregated compliance	System, inter & intra	Observational			•	
4.5b. Sentence level compliance	System, inter & intra	Observational	•		•	
4.6a. Fixed effects	Inter	Hierarchical		•	•	
4.6b. Random intercepts	Inter	Hierarchical		•	•	
4.6c. Random slopes	Inter	Hierarchical	•	•	•	

On the positive side, given the simplicity of the research design we can consider experimental simulations a useful tool to generate inter-judge comparisons across time and jurisdictions for specific types of offences.

4.2. Randomised caseloads

The use of randomisation for the study of consistency in sentencing was first employed over 80 years ago by Gaudet et al. (1933). This method relies on the practices followed by most federal courts in the US where judges in the same location are assigned cases randomly to prevent 'judge shopping' and help ensure fair procedures. This random process guarantees that over a large number of cases, each judge in a court area will be assigned a similar mix of cases.

We can use this type of data to obtain measures of intra-judge consistency for individual judges by assessing how disperse their sentence outcomes are compared to the average dispersion across judges. Perhaps more interestingly, we can also use this type of data, to obtain a parsimonious measure of inter-judge consistency by taking the average of the sentence outcomes passed by each individual judge and assessing their dispersion.[27] This measure of consistency is appealing for its simplicity, as demonstrated inWaldfogel (1991), Orchard et al. (1997) and Scott (2010).[28]

[27]Taking $\bar{\bar{l}}$ to represent the jurisdiction's mean sentence outcome, we can formally express this measure of consistency as follows, $\widehat{C} = \sqrt{1/J - 1\sum_{j=1}^{J}(\bar{l}_j - \bar{\bar{l}})^2}$.

[28]A more sophisticated approach, applied in these same three studies, involves the implementation of an ANOVA test. Such an approach tests whether the variability in the average sentences amongst judges is greater than would be expected from the variability created by the random allocation of cases. If the ANOVA shows that there are statistically significant differences between judges, then there are a number of statistics that can be used to calculate the size of this effect – for instance omega squared, partial eta, or their generalised counterparts (Olejnik and Algina, 2003). Alternatively, other methods that have been used in the literature when randomised case-loads are available involve the specification of fixed or random effects models. In its simplest form, the fixed effect model takes the following form: $Y_l = \beta_0 + \beta_j M_{jl} + u_l$, where Y_l captures the length of each sentence, M_{jl} represents a set of dummy variables for the different judges, the regression coefficients are denoted by β, and u_l represents the error term. A joint test of the coefficients on the judge dummy variables provides a valid test of whether the difference between judges is statistically significant.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

The use of real sentences improves the validity of this method compared to *experimental simulations*, while the use of randomly allocated cases removes the need to rely on statistical controls to isolate $Var(I)$ from $Var(L)$. The importance of this last feature cannot be overemphasised, it renders the debate of what is to be considered a legal factor irrelevant and eliminates typical misspecification problems (i.e. omitted relevant variables, multicollinearity, measurement error, etc.) associated with statistical modelling techniques.

Perhaps the main disadvantage of *randomised caseloads* stems from its incapacity to generate measures of system consistency. Using measures of inter-judge consistency as a proxy for the overall level of consistency can still be quite informative. However, the validity of these measures should also be called into question. The measure of inter-judge consistency obtained from this method would not be able to capture treatments of different cases that are 'compensated' across judges. For example, Hofer (2007) presented a hypothetical case where some judges treat drugs more harshly than fraud with others doing the opposite to illustrate this point.

The presence of such problem will produce an upward biased estimate of consistency, although this could be offset by another potential problem that relates to the applicability of the method, which could produce a bias on the opposite direction. The randomised case allocation process only guarantees that different judges will receive the same average caseload over a very large number of cases. In real-world sample sizes, differences will exist in the average seriousness of the caseload assigned to different judges, purely as a result of the random assignment process. The dispersion of judge's averages could be a result of the mix of cases they happened to receive rather than a genuine difference in their sentencing practice.

In addition we could note two further limitations regarding the generalisability and comparability of findings produced from randomised caseloads. In its simplest form this method can only provide comparison of judges within courts, and as a result it cannot be used to generalise to an entire jurisdiction. The problem stems from the fact that case randomisation occurs only within courts, meaning that there is no guarantee the caseload of judges in different courts will be similar, even over a large number of cases.[29] Regarding the comparability of findings across jurisdictions we need to note that the policy of randomising caseloads is not common, and only a few of the US districts following this practice have made their sentencing data available for research. Finally, given the big samples that are required per judge to make the randomisation effective, it should not be expected to allow time comparisons in a scale smaller than a year.[30]

4.3. (Un)conditional comparisons & exact matching

Although experimental data is scarce almost all OECD jurisdictions make available observational data covering at least sentence outcomes and type of offences. Using this data we can obtain a blunt assessment of consistency by comparing the variability of sentence outcome conditional on the type of offence. Some

[29] To overcome this problem, Hofer et al. (1999) and Scott (2010) extend the design to include court level dummies as follows: $Y_l = \beta_0 + \beta_r D_{rl} + \beta_j M_{jl} + u_l$, where D_{rl} represents the dummy court variables, indexed by $r = 1, 2, \ldots, R$. A measure of inconsistency can thus be obtained by comparing the adjusted R^2 of this model to that of a simplified version in which the judge dummies are excluded. The downside of this approach is that it cannot account for correlations in sentencing practice within courts, which will be present if judges' sentencing practice is influenced by that of their peers within the same court. If some courts tend to be more lenient than others, then the model will not be able to detect this type of inconsistency. A similar approach, which shares the same disadvantage, involves the specification of a random effects model to detect differences between judges. For example, Anderson et al. (1999) attributed differences between judges using a random intercept term instead of the set of dummy judge variables. This model requires ordering the data under a hierarchical structure of sentences within judges and breaking the residual term in two parts: e_j, capturing the sentence level unexplained variability, and τ_j, for that at the judge level, $Y_{lj} = \beta_0 + \beta_r D_{rlj} + e_l + \tau_j$. A key advantage of the random effects model is it allows the calculation of 95% confidence levels for the random intercept term, which provides a measure of the size of the difference in sentencing practice between different judges.

[30] For example, a simple simulation we ran shows that with a sample of 300 sentences drawn from a normal distribution with a standard deviation of five years, only differences of half a year or more in the average sentence length between two judges can be detected with a 5% significance level.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

examples of studies using this design are Lovegrove (1984) and Walker and Sager (1991). When hierarchical data is available a similar design can be run by comparing the mean sentence outcome by court. Tarling (2006) used both designs to compare dispersion in disposal types amongst offences of burglary in 1974 and 2000, and differences in disposal type between 30 magistrates' courts of England and Wales between the same years. Mason et al. (2007) combined these two approaches by analysing sentence length variability between magistrates and Crown Courts controlling for different variables such as type of offence or local crime rates.

These methodologies are limited in their ability to control simultaneously for multiple confounding effects such as type and severity of offences, and the socio-demographic characteristics of different courts. As a result, observed disparities in sentence outcomes may be due to differences between caseloads and do not necessarily imply inconsistent sentencing practices. This approach can however be improved when additional data describing the legal factors defining the case is available using an *exact matching* design. This was first proposed by Hofer et al. (1999), who suggested that to address this comparability problem we can use 'matched-groups' of offences. These are groups of offences that share the same characteristics, so the variability of sentence outcomes within each of these groups can be attributed to inconsistencies in sentencing.

The use of controls improves the validity of *exact matching* compared to the *conditional comparison* of variances. However, in practice, it is difficult – perhaps impossible – to find a sentencing dataset where all the relevant legal factors are covered. So long as one of them is missing the final measure of dispersion obtained from this method will capture part of $Var(L)$, and therefore will represent a biased (upwards) approximation to the real level of inconsistency. Furthermore, even if all the relevant legal factors were available to the researcher, Hofer et al. (1999) pointed at a technical limitation that studies using *exact matching* will ultimately have to face. The more controls used to define the matched groups the more groups will be generated, which reduces the available sample size of each group, and therefore the statistical power to detect significant variations. That is, there is a trade-off between the validity of the measure of consistency and the reliability of that measure (i.e. there is a trade-off between bias and precision).[31]

Furthermore, in its simplest form, *conditional comparisons* and *exact matching* create measures of consistency for specific types of offences, which might not be representative of the sentencing process in the entire jurisdiction. The external validity of the method could be enhanced by taking a sample of different types of offences (or matched groups) weighting the results from each of them based on their relative frequency, and then aggregating all these weighted estimates into a single measure (Pina-Sánchez and Linacre, 2014).

In spite of these limitations *exact matching* can still offer very informative findings. It was used by the US Sentencing Commission (1991) in its evaluation of sentencing guidelines, where offences of bank robbery and cocaine/heroin distribution were studied separately. To match drug offences the following controls were used: the amount of drugs, injury caused to any victims, the defendant's role on the offence, criminal record, and whether the defendant pleaded guilty. These groups were then used to compare variances in the sentence outcomes within 'matched groups' before and after the guidelines came into force.[32,33]

Notice how these measures of consistency are bound to be biased since the matched groups are missing several legal factors to be considered when sentencing drug offences. However, we could still legitimately use them to assess changes in consistency across time if the same legal factors are omitted in each period.

[31]This problem is akin to what Bellman (1961) defined as the curse of dimensionality.

[32]When the interest lies in assessing changes across time, the different group variances at two different points can be tested using an F test for the ratio of variances, such as: $F = \sigma_B^2/\sigma_A^2$, where σ_B^2 and σ_A^2 represent the before and after variances, and the degrees of freedom are given by the sample size of the respective variances.

[33]To generate before and after weighted comparisons we could use the following statistic, $\sum_{k=1}^{K}(n_k/N(\sigma_B^2/\sigma_A^2)/\overline{W})$, where $k = 1, 2, \ldots, K$ is used to index the different matched groups and the mean weight, $\overline{W}$, is used so the group weights can have a mean of one while the scale of the variances remains unchanged.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

That is, we will not be able to obtain a general estimate of consistency but a relative one with which we could still carry out comparisons across time.

On the other hand, as it was the case for randomised caseloads, the requirement of using big samples to make *exact matching* effective rules out the possibility of observing changes across time in a smaller scale than annually. Such a research question can be better explored using the following method.

4.4. Dispersion of residuals

The purpose of controlling for multiple confounding effects can be more efficiently achieved using regression analysis than *exact matching*. Specifically, we can use regression techniques to model sentence outcomes based on relevant legal factors. Unlike *exact matching*, this method allows assessing multiple types of offences simultaneously, without having to divide the sample size after a new legal factor is included, which improves its generalisability. Consistency is approximated using a measure of the goodness of fit of the model,[34] while a measure of inconsistency could be obtained from the variability in sentence outcomes that is unexplained by the model, i.e. the model's residuals.

It is widely recognised that this design – to be denoted here as *cross-sectional analysis of residuals* – is flawed (Brantingham, 1985; Waldfogel, 1998; Anderson et al., 1999; Hofer et al., 1999). As indicated before, it would be impossible to control for all the relevant legal factors, hence, some of the unexplained variability in sentencing would be wrongly attributed to inconsistency in sentencing. In addition, and in contrast with the *exact matching* approach, regression models assume a specific functional form, which could result in problems of misspecification. Hofer et al. (1999) are particularly critical of this approach: "*these studies can add relatively little, if anything, to our understanding of disparity under the guidelines, and are even less helpful in evaluating whether the amount of disparity has increased or decreased*" (p. 244).

We oppose this view, and contend that although regression techniques are not useful for assessing consistency at a point in time, careful application can yield useful insights into how consistency is changing through time, much like we noted for the case of *exact matching*. If it can be assumed that missing relevant legal factors and model misspecifications will remain constant through time, then changes in the variability of residuals will be a valid measure of changes in consistency, even though the exact level at any particular time is unknown (Ostrom et al., 2008; Pina-Sánchez and Linacre, 2014). We denote this as the *longitudinal analysis of residuals*.

One criticism made by Hofer et al. (1999) – with which we agree – relates to the limitations of using measures of discrete changes in time like comparisons of the goodness of fit of two different models, which "*cannot detect trends that were occurring before implementation, and the results may mistakenly be taken to suggest that all differences between the two times are due to the guidelines*" (p. 267). As a solution to this problem Pina-Sánchez and Linacre (2014), suggested deriving a pattern of change by running a single model and generating measures of unexplained variability by week or month.[35]

The possibility of observing changes in consistency on a continuous scale makes this design an ideal exploratory tool to carry out across time comparisons. For example, Pina-Sánchez and Linacre (2014) found a trend towards the reduction in dispersion in England and Wales across 2011, which did not correspond to a structural change immediately after a new guideline came into force.

4.5. Guidelines and sentence level compliance

When a system of sentencing guidelines is present, or more generally, when the appropriate sentence outcome for specific types of offences is well defined, we can use the recommended sentence outcomes as

[34] Such as the R^2.

[35] Specifically the authors used the weekly residual variance, which can be expressed as, $\sum_{l=1}^{N_w}(\mu_{lw}-\overline{\mu})^2/N_w$, where $\overline{\mu}$ represents the mean of residuals for the whole year (by assumption equal to zero), w is a subscript indicating the weeks of the year so $w = 1, 2, \ldots, 53$, and N_w represents the sample size in one particular week.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

benchmarks to carry out normative analyses of consistency.[36] For example, a simple measure of consistency can be generated from the proportion of cases falling within the recommended outcome(s). These 'compliance statistics' are routinely published by all US sentencing commissions and independent researchers (Minnesota Sentencing Guidelines Commission, 2012; Oregon Criminal Sentencing Commission, 2003; Frase, 2005b; Hofer, 2007; Kramer and Ulmer, 2002; Roberts, 2013b; Scott, 2010; Tonry, 1987; Ulmer et al., 2011).

Analyses of compliance have mainly been carried out for specific categories of offences. To increase the generalisability of the method and obtain measures of system consistency we could extend the analysis to every type of offence. This solution could however represent a very laborious task since for each offence different benchmarks are used, instead we suggest taking a sample of some of those offences and weighting them based on their relative frequency. The degree of comparability of these findings across time is quite straightforward, while comparisons across jurisdictions are limited to jurisdictions governed by specific guidelines.

Perhaps the biggest weakness of the method stems from its questionable validity. These aggregated measures of compliance only offer a dichotomous view into the problem of consistency. They look at whether sentences fall within the normative bands or not. They do not take into account how far or close sentences fall from the recommended outcome, or the extent to which departures were justified.

This design can be improved by calculating the specific sentence that should have been passed according to: 1) the legal factors defining a particular case, and 2) what is specified in the guidelines. The difference between this normative point and the actual sentence could be taken as evidence of inconsistency.[37] Examples of studies approximating this design include: Scott (2010), where the author looked at the judges' average sentencing distance from the US guideline range, and Waldfogel (1998) who determined the most appropriate sentence as the average sentence for an offender with given circumstances, and assessed inconsistency by taking squared deviation between each of these sentences and their mean.

Another group of studies have made use of the − known or estimated − presumptive sentence by including it as a factor modelling different sentence outcomes (Bushway and Piehl, 2001; Engen and Gainey, 2000; Griswold, 1987; Mason and Bjerk, 2013). We could take the goodness of fit of such a model or the unexplained variability as measures of (in)consistency, just like for the *dispersion of residuals*. Interestingly, if the presumptive sentence is known this design eliminates the limitation of having to control for all of the relevant legal factors of the case. Otherwise, it will be prone to same problems mentioned above.

4.6. Fixed and random effects models

We can overcome many of the limitations affecting designs relying on observational data when a hierarchical structure is present in the dataset. If we have identifiers of judges or courts we can include them in a regression model of sentence outcomes (a fixed effects model) and test whether they have a statistically significant effect. The greater the effect of these judge and/or court identifiers, the more evidence there is of inconsistent sentencing.

Waldfogel (1998) and Pina-Sánchez and Linacre (2013) used this design as part of their exploratory analysis. This method shares the same features of generalisability and comparability as the *dispersion of residuals*; similarly, is also affected by misspecifications, in particular in the form of omitted relevant legal factors. However, we argue that the validity of the coefficients capturing the judge and/or court effect as a measure of consistency is higher than what is captured by the goodness of fit or by the residuals from a model including just legal factors. These regression coefficients for the judges/courts might be biased but

[36] "*The principal goal of all guidelines schemes is to promote consistency, and the effectiveness of any sentencing guideline scheme in achieving this objective may be measured by the proportion of sentences falling within the guidelines. This statistic is in large measure determined by the nature of the compliance requirement upon courts*" (Roberts, 2011, p. 997).

[37] Such measure of consistency could be expressed as, $\widehat{C} = |l - l^*|/N$, where l^* denotes the normative point.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

they capture essentially unexplained court differences in sentencing, whereas the residuals from a model using legal factors will confound that with any other possible source of unexplained variance.

A conceptually similar measure of inconsistency can also be obtained using a *random intercepts* model. This is nothing more than a regression model with an intercept that is allowed to vary to reflect differences at the judge or court level that cannot be explained by the model. Interestingly, we can compare this unexplained variability stemming from the higher level (judges or courts) over the overall unexplained variability in the model. This is known as the intra-cluster correlation (ICC), and it has been used in studies such Fearn (2005), Ulmer et al. (2011) and Pina-Sánchez and Linacre (2013) to obtain additional insights into consistency.

The *random intercepts* model can be extended to include *random slopes* for some of the legal factors used as explanatory variables.[38] This extension removes the simplifying assumption that legal factors must have a constant effect on sentences across courts. In so doing it allows sentences to be broken down into the different legal factors that compose them in order to detect the main factors responsible for the observed inconsistencies. This specific measures of consistency can be extremely useful for the policy maker; however, notice that it doesn't offer a measure of system consistency, but one related to the application of specific legal factors.

Anderson and Spohn (2010) used *random slopes* models to assess inter-judge variability in three US District Courts after the implementation of the Federal Sentencing Guidelines, and Pina-Sánchez and Linacre (2013) replicated this analysis for offences of assault sentenced in the Crown Court of England and Wales. A key advantage to the analysis of random slopes is that the technique may be more robust to problems of omitted variables, and so may be more suitable for use with observational data which do not include comprehensive information about all the relevant legal factors. In particular, for a random slope to be biased by omitted relevant variables, we would need not only that those omitted variables are associated with the factors included in the model, but also that their effect differed across courts.[39]

5. Conclusion

The promotion of consistency in sentencing has become a goal of many reforms of the sentencing process across the world. Yet very little is known about the actual levels of consistency in sentencing across time or jurisdictions. Crucial research questions such as the level of consistency in magistrates' courts of the UK, or how this compares to the District Courts in the US, remain unanswered. This relative state of ignorance is surprising and it needs to be overcome if we want future reforms to be based on facts.

We have argued that to a certain extent the gulf between the importance and lack of evidence is due to conceptual and methodological confusion as to what is consistency and how can we measure it. In particular, after reviewing the literature we came to the conclusion that a multiplicity of methods have been used to study consistency in sentencing, many of them are based on different definitions consistency, and almost none of them offer findings of similar scope or robustness, making comparisons between studies very unreliable. In the attempt to improve this situation we have produced in this article a theoretical and methodological review of the concept of consistency. Specifically, we have done the following:

1) We refined the definition of consistency in sentencing by deconstructing what is meant by 'like cases' 'treated alike' and by contrasting it to some other concepts with which it is often confounded (proportionality, uniformity and discrimination).

[38] More formally, this model extends the random effects model presented in footnote 18 by including the *random slope* terms, ϑ_{ij}, for those legal factors of interest, $Y_{lj} = \beta_0 + (\beta_l + \vartheta_{lj})D_{rlj} + e_l + \tau_j$.

[39] Further discussed in Pina-Sánchez and Linacre (2013, pp. 6–7).

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

2) We spelt out the challenges of operationalising the concept of consistency empirically (the isolation of legitimate from illegitimate variability) and identified the different elements of the concept that can be quantified (system, inter and intra-judge consistency).
3) We reviewed and categorised eleven methods that have been used in the literature to measure outcome consistency in sentencing according to their scope, type of data required, the validity of their assumptions, and whether they produce generalisable and comparable findings.

A number of key points emerged from this review. First, *randomised case-loads*, or the methods based on randomly allocated sentences, have often been considered as the gold standard method to measure consistency. However, we have noted that these research designs fail to grasp intra-judge sources of inconsistency and cannot be adequately used to make comparisons across time or jurisdictions.

Second, methods relying on observational data such as *exact matching* and the *analysis of residuals* can only provide biased measures of consistency. However, they can look at both inter and intra-judge sources of inconsistency simultaneously and they provide an interesting tool to depict changes in consistency through time, which may be less biased.

Third, in the presence of hierarchical data, we can explore further avenues of research for the measurement of consistency using multilevel models. When detailed information on the legal factors present in each case is also available, the implementation of *random intercepts* models may provide more robust measures of inconsistency between judges or courts than an analysis of the residuals using a similar but non-hierarchical model. Furthermore, we can also use *random slopes* models to detect the legal factors more inconsistently applied.

More generally, this review has also pointed out at the wide differences in scope and validity offered by different methods, from which we conclude that to obtain a reliable insight into the overall level of consistency in a jurisdiction we need to produce evidence from more than one method. Where only one method is used, it would be important that the authors refer to the elements of consistency that are being measured and the limitations to the validity of the method. We believe that the classification offered in this paper will facilitate the articulation of such distinctions more clearly, and in so doing – hopefully – improve the study of consistency in sentencing.

References

Albonetti, C.A., 1997. Sentencing under the federal guidelines: effects of defendant characteristics, guilty pleas, and departures on sentence outcomes for drug offenses, 1991–1992. Law Soc. Rev. 789–822.

Albonetti, C.A., 1998. Direct and indirect effects of case complexity, guilty pleas, and offender characteristics on sentencing for offenders convicted of a white-collar offense prior to sentencing guidelines. J. Quant. Criminol. 14, 353–378.

Albonetti, C.A., 2002. The joint conditioning effects of defendant's gender and ethnicity on length of imprisonment under the federal sentencing guidelines for drug trafficking/manufacturing offenders. J. Gend. Race Justice 6, 39–60.

Alschuler, A.W., 2005. Disparity: the normative and empirical failure of the federal guidelines. Stanf. Law Rev. 85–117.

Anderson, J., Kling, J., Stith, K., 1999. Measuring inter-judge sentencing disparity: before and after the federal sentencing guidelines. J. Law Econ. 42, 271–307.

Anderson, A., Spohn, C., 2010. Lawlessness in the federal sentencing process: a test for uniformity and consistency in sentence outcomes. Justice Q. 27, 362–393.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

Austin, J., Jones, C., Kramer, J., Renninger, P., 1996. National assessment of structured sentencing. Bureau of Justice Assistant, Washington, DC.

Baumer, E.P., 2013. Reassessing and redirecting research on race and sentencing. Justice Q. 30, 231–261.

Bellman, R.E., 1961. Adaptive Control Processes: a Guided Tour. University Press, Princeton.

Brantingham, P., 1985. Sentencing disparity: an analysis of judicial consistency. J. Quant. Criminol. 1, 281–305.

Britt, C.L., 2009. Modeling the distribution of sentence length decisions under a guidelines system: an application of quantile regression models. J. Quant. Criminol. 24, 341–370.

Bushway, S.D., Piehl, A.M., 2001. Judging judicial discretion: legal factors and racial discrimination in sentencing. Law Soc. Rev. 733–764.

Casey, J., Wilson, J., 1998. Discretion, disparity or discrepancy? A review of sentencing consistency. Psychiatry Psychol. Law 5, 237–247.

Cole, K., 1997. The empty idea of sentencing disparity. Northwest. Univ. Law Rev. 91.

Davies, M., Takala, J.P., Tyrer, J., 2002. Sentencing burglars in England and Finland: a pilot study. In: Tata, C., Hutton, N. (Eds.), Sentencing and Society: International Perspectives. Ashgate, Aldershot, pp. 257–276.

Davies, M., Tyrer, J., 2003. 'Filling in the gaps' a study of judicial culture: views of judges in England and Wales on sentencing domestic burglars contrasted with the recommendations of the Sentencing Advisory Panel and the Court of Appeal guidelines. Crim. Law Rev. 243–265.

Danziger, S., Levav, J., Avnaim-Pesso, L., 2011. Extraneous factors in judicial decisions. Proc. Natl. Acad. Sci. 108, 6889–6892.

Dhami, M.K., 2013. A 'decision science' perspective on the old and new sentencing guidelines in England and Wales. In: Ashworth, A., Roberts, J.V. (Eds.), Structured Sentencing in England and Wales: from Guidance to Guidelines. University Press, Oxford, pp. 165–181.

Engen, R.L., 2011. Racial disparity in the wake of Booker/Fanfan: making sense of "messy" results and other challenges of sentencing research. Criminol. Public Policy 10, 1139–1149.

Engen, R.L., Gainey, R.R., 2000. Modeling the effects of legally relevant and extralegal factors under sentencing guidelines: the rules have changed. Criminology 38, 1207–1230.

Everett, R.S., Wojtkiewicz, A., 2002. Difference, disparity, and race/ethnic bias in federal sentencing. J. Quant. Criminol. 18, 189–211.

Fearn, N.E., 2005. A multilevel analysis of community effects on criminal sentencing. Justice Q. 22, 452–487.

Fischman, J.B., Schanzenbach, M.M., 2012. Racial disparities under the federal sentencing guidelines: the role of judicial discretion and mandatory minimums. J. Empir. Leg. Stud. 9, 729–764.

Frankel, M., 1972. Lawlessness in sentencing. Univ. Cincinnati Law Rev. 41, 1–54.

Frase, R.S., 2005a. State sentencing guidelines: diversity, consensus, and unresolved policy issues. Columbia Law Rev. 105, 1190–1232.

Frase, R.S., 2005b. Sentencing guidelines in Minnesota, 1978–2003. Crime Justice 32, 131–219.

Gaudet, F., Harris, G., St John, Ch, 1933. Individual differences in the sentencing tendencies of judges. J. Crim. Law Criminol. 23, 191–208.

Griswold, D.B., 1987. Deviation from sentencing guidelines: the issue of unwarranted disparity. J. Crim. Justice 15, 317–329.

Hofer, P.J., Blackwell, K., Ruback, R.B., 1999. The effect of the federal sentencing guidelines on inter-judge sentencing disparity. J. Crim. Law Criminol. 90, 239–321.

Hofer, P.J., 2007. United States vs booker as a natural Experiment: using empirical research to inform the federal sentencing policy debate. Criminol. Public Policy 6, 433–460.

Hola, B., 2012. Sentencing of international crimes: consistency of case law. Amst. Law Forum 3–24.

Hough, M., Jacobson, J., Millie, A., 2003. The Decision to Imprison: Sentencing and the Prison Population. Prison Reform Trust, London.

Kramer, J., Ulmer, J., 2002. Downward departures for serious violent offenders: local court 'corrections' to Pennsylvania's sentencing guidelines. Criminology 40, 897–932.

Krasnostein, S., Freiberg, A., 2013. Pursuing consistency in a individualistic sentencing framework: if you know where you're going, how do you know when you've got there. Law Contemp. Probl. 76, 265–288.

Lovegrove, A., 1984. An empirical study of sentencing disparity among judges in an Australian criminal court. Int. Rev. Appl. Psychol. 33, 161–176.

Mason, C.E., Bjerk, D., 2013. Inter-judge sentencing disparity on the federal bench: an examination of drug smuggling cases in the southern district of California. Fed. Sentencing Report. 25, 190–196.

Mason, T., de Silva, N., Sharma, N., Brown, D., Harper, G., 2007. Local Variation in Sentencing in England and Wales. Ministry of Justice, London.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

Minnesota Sentencing Guidelines Commission, 2012. Sentencing Practices: Controlled Substance Offenses Sentenced in 2010 from. http://www.leg.state.mn.us/docs/2013/other/130860.pdf.
Minnesota Sentencing Guidelines Commission, 2014. Minnesota Sentencing Guidelines and Commentary from. http://mn.gov/sentencing-guidelines/images/2014%2520Guidelines.pdf.
Mustard, D.B., 2001. Racial ethnic and gender disparities in sentencing: evidence from the US federal courts. J. Law Econ. 44, 285–314.
Olejnik, S., Algina, J., 2003. Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol. Methods 8, 434.
Orchard, N., Howlett, J., Davies, E., Pearson, G., Payne, A., 1997. Does inter judge disparity really matter? An analysis of the effects of sentencing reforms in three federal district courts. Int. Rev. Law Econ. 17, 337–366.
Oregon Criminal Sentencing Commission, 2003. Sentencing practices: Summary Statistics for Felony Offenders Sentenced in 2001 from. http://www.oregon.gov/CJC/docs/SG01v2.pdf.
Ostrom, B., Ostrom, Ch, Hanson, R., Kleiman, M., 2008. Assessing Consistency and Fairness in Sentencing: a Comparative Study in Three States. National Institute of Justice, Washington.
Pasko, L., 2002. Villain or victim: regional variation and ethnic disparity in federal drug offense sentencing. Crim. Justice Policy Rev. 13, 307–328.
Peterson, R.D., Hagan, J., 1984. Changing conceptions of race: towards an account of sentencing research. Am. Sociol. Rev. 56–70.
Pina-Sánchez, J., Linacre, R., 2013. Sentence consistency in England and Wales: evidence from the crown court sentencing survey. Br. J. Criminol. 53, 1118–1138.
Pina-Sánchez, J., Linacre, R., 2014. Enhancing consistency in sentencing: exploring the effects of guidelines in England and Wales. J. Quant. Criminol. 30, 731–748.
Reitz, K.R., 2013. Comparing sentencing guidelines: do us systems have anything worthwhile to offer England and Wales? In: Ashworth, A., Roberts, J. (Eds.), Sentencing Guidelines: Perspectives on the Definitive Guidelines. Oxford University Press, Oxford, pp. 182–201.
Roberts, J.V., 2011. Sentencing guidelines and judicial discretion: evolution of the duty of courts to comply in England and Wales. Br. J. Criminol. 51, 997–1013.
Roberts, J.V., 2013a. Sentencing guidelines in England and Wales: recent developments and emerging issues. Law Contemp. Probl. 76, 1–25.
Roberts, J.V., 2013b. Complying with sentencing guidelines: latest findings from the Crown Court Sentencing Survey. In: Ashworth, A., Roberts, J. (Eds.), Sentencing Guidelines: Perspectives on the Definitive Guidelines. Oxford University Press, Oxford, pp. 104–120.
Schanzenbach, M.M., Tiller, E.H., 2008. Reviewing the sentencing guidelines: Judicial politics, empirical evidence, and reform. The University of Chicago Law Review 75, 715–760.
Scott, R., 2010. Inter-judge sentencing disparity after booker: a first look. Stanf. Law Rev. 63 from. http://www.stanfordlawreview.org/sites/default/files/articles/Scott_63_Stan._L._Rev._1_0.pdf.
Sebba, L., 2013. Is sentencing reform a lost cause: a historical perspective on conceptual problems in sentencing research. Law Contemp. Probl. 76, 237–264.
Sentencing Council, 2011. Analytical Note: the Resource Effects of Increased Consistency in Sentencing. Sentencing Council Documents, from. http://sentencingcouncil.judiciary.gov.uk/docs/Consistency_in_sentencing.pdf.
Spohn, C., 2000. Thirty years of sentencing reform: the quest for a racially neutral sentencing process. Crim. Justice 3, 427–501.
Spohn, C., 2002. Sentencing: Disparity. Encyclopedia of Crime and Justice from. http://www.encyclopedia.com/doc/1G2-3403000240.html.
Stacey, A.M., Spohn, C., 2006. Gender and the social costs of sentencing: an analysis of sentences imposed on male and female offenders in three US district courts. Berkeley J. Crim. Law 11, 43–76.
Starr, S.B., Rehavi, M.M., 2013. Mandatory sentencing and racial disparity: assessing the role of prosecutors and the effects of booker. Yale Law J. 123, 2–80.
Steffensmeier, D., Demuth, S., 2000. Ethnicity and sentencing outcomes in US federal courts: who is punished more harshly? Am. Sociol. Rev. 705–729.
Stolzenberg, L., D'Alessio, J., 1994. Sentencing and unwarranted disparity: an empirical assessment of the long-term impact of sentencing guidelines in Minnesota. Criminology 32, 301–310.
Tarling, R., 2006. Sentencing Practice in Magistrates' Courts Revisited. Howard Journal 45, 29–41.
Tonry, M.H., 1987. Sentencing Reform Impacts. National Institute of Justice, Rockville.
Tonry, M.H., 1996. Sentencing Matters. Oxford University Press, New York.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001

Ulmer, J., Light, M., Kramer, J., 2011. The 'liberation' of federal judges' discretion in the wake of the Booker/Fanfan decision: is there increased disparity and divergence between courts? Justice Q. 28, 799–837.

United States Sentencing Commission, 1991. The federal sentencing guidelines: A report on the operation of the guidelines system short term impacts on disparity in sentencing, use of incarceration, and prosecutorial discretion and plea bargaining. USSC, Washington, DC.

United States Sentencing Commission, 2004. Fifteen years of guidelines sentencing: An assessment of how well the federal criminal justice system is achieving the goals of sentencing reform. USSC, Washington, DC.

Wandall, R.H., 2006. Equality by numbers or words: a comparative study of sentencing structures in Minnesota and in Denmark. Crim. Law Forum 17, 1–41.

Waldfogel, J., 1991. Aggregate inter-judge disparity in federal sentencing: evidence from three districts. Fed. Sentencing Report. 4, 151–154.

Waldfogel, J., 1998. Does inter-judge disparity justify empirically based sentencing guidelines? Int. Rev. Law Econ. 18, 293–304.

Walker, T., Sager, T., 1991. Are the federal sentencing guidelines meeting congressional goals?: an empirical and case law analysis. Emory Law J. 40, 393–444.

Please cite this article in press as: Pina-Sánchez, J., Linacre, R., Refining the measurement of consistency in sentencing: A methodological review, International Journal of Law, Crime and Justice (2015), http://dx.doi.org/10.1016/j.ijlcj.2015.06.001