Do sentencing guidelines result in lower inter-judge disparity ? Evidence from framed field experiment

Cécile Bourreau-Dubois, Myriam Doriat-Duban, Bruno Jeandidier, Jean-Claude Ray

BETA
Bureau
d'économie
théorique
et appliquée

Documents de travail

« Do sentencing guidelines result in lower inter-judge disparity? Evidence from a framed field experiment »

Auteurs

Cécile Bourreau-Dubois, Myriam Doriat-Duban,
Bruno Jeandidier, Jean-Claude Ray

Document de Travail n° 2021 – 17
(Version révisée du WP 2020-28)

Mai 2021

Bureau d'Économie Théorique et Appliquée
BETA

www.beta-umr7522.fr

@beta_economics

Contact :
jaoulgrammare@beta-cnrs.unistra.fr

Université de Strasbourg · cnrs · Université de Lorraine · INRAE · AgroParisTech

Do sentencing guidelines result in lower inter-judge disparity? Evidence from a framed field experiment[1]

Cécile Bourreau-Dubois, Myriam Doriat-Duban, Bruno Jeandidier, Jean-Claude Ray

Université de Lorraine, Université de Strasbourg, CNRS, BETA, F-54000 Nancy, France

Mai 2021

Abstract: We study the decision-making of judges in an experimental setting resembling real world judicial decision-making. We gave 312 future judges 48 vignettes built from real data related to divorce cases involving children. We compared two different subject pools: judges who were asked to set child support awards with an advisory guideline and judges who were asked to set child support awards without any guidelines. We found that the introduction of such a guideline helps to reduce the disparity between judges (i.e., the variance in similar cases is lower when the subjects have the opportunity to use the guideline) but that this effect is not systematic, as an increase in heterogeneity was observed in some specific cases.

Keywords: controlled experiment - field experiment - judicial sentencing - child support - guidelines

JEL code: K42

[1] This work has benefited from the financial support of the Mission Droit et Justice (2016-2018).

Do sentencing guidelines result in lower inter-judge disparity? Evidence from a framed field experiment

Introduction

The judicial systems of developed countries have been marked over the last 30 years by the development of guidelines. In the United States, two areas have been affected by this since the mid-1980s: the length of prison sentences handed down to criminal defendants and the amount of child support awards. In France, two official guidelines were implemented more recently: an advisory child support guideline in 2010 and a mandatory guideline for setting damages in labor courts in 2017. One of the major goals of the guidelines is to improve horizontal equity, through the reduction of unwarranted sentencing disparity, that is, a different treatment for similarly situated cases (Waldfogel, 1998).[2] Such disparity may have different origins. First, a decision maker may set different sentences for similar cases from the judicial point of view. This may be observed if the judge is influenced in her decision by factors such as the political, economic or social context of the case (Ichino et al., 2003; Marinescu, 2011), the qualitative characteristics of the case (e.g., the quality of the lawyer, the remorse of the offender) or even the personal characteristics of the litigants (age, gender or ethnic origin). Furthermore, judges, like other economic agents, may also be considered victims of cognitive biases and errors (Guthrie et al., 2001; Wistrich et al., 2015; Spamann and Klöhn, 2016; Liu, 2018; Kahan, 2015). In such cases, the relative bias of their decisions would result from heuristics such as confirmation, anchoring or availability. As a result, a judge's sentencing may be different for similarly situated litigants. The second origin of disparity is between-judge variation. Judges may be more or less influenced by factors that are not in the law (whether legitimately or not) (Abrams et al., 2012). The individual characteristics of the judge may also impact on her decisions, such as her gender (Peresie, 2005), ethnic origin (Abrams et al., 2012) or political affiliation (Cohen and Yang, 2019). Even if judges are not driven by these kinds of factors, different judges could use different sentencing for similar cases because they may give greater or lesser weight to legally relevant factors due to their individual preferences (Waldfogel, 1998; Woolredge, 2010). As a result of these philosophical and attitudinal differences, the sentence may rest in part on the judge who issues it. In this paper we focus on this type of disparity.

Guidelines, in particular when they are mandatory, are expected to reduce the inter-judge disparity resulting from the philosophy, ideology or bias of the sentencing judge, by constraining judicial

[2] The elimination of unwarranted disparity is not the only goal of the guidelines. For instance, the US guidelines had other objectives such as promoting deterrence by increasing the length of prison sentence prisons or reducing poverty among children by increasing the amount of child support.

discretion. Numerous studies have explored whether in practice these guidelines have been successful in achieving a reduction of disparity in decision-making. These studies mobilize real databases (mainly resulting from the sentencing activity of courts), allowing for a comparison of decisions before and after the introduction of the guidelines. One limitation of this approach is that they make it difficult to separate the pure effects of the guidelines from the effects related to characteristics that are observed by the judges during the judgment but not by the researcher. Thus, there is a risk of biased estimates due to unobserved heterogeneity. In most studies this problem is considered as solved provided that cases are assigned randomly to judges, ensuring that judges receive the same distribution of case characteristics, both observed and unobserved. A second limitation of these studies is that while they can examine overall inter-judge heterogeneity in sentencing, they have more difficulty identifying the combination of characteristics which lead to a higher risk of sentencing disparity. Consequently, one may consider that the study of inter-judge sentencing disparity on the basis of real data does not make it possible to assess the effectiveness of a guideline in terms of reducing disparity in a fully convincing manner.

To circumvent this limitation, we used a controlled experiment. Our subjects are people who have passed the exam to become judges and who, after three years spent at the Ecole Nationale de la Magistrature (ENM), will be appointed as judges in French courts. They were tasked with setting child support amounts for the same given cases, each one characterized by a limited amount of information: the income of the parties, the support amounts proposed by the parties, the age and mode of residence of the child, and the number of children of the couple. The treatment consisted in enabling some of the subjects to set the child support using the child support guidelines made available by the French Ministry of Justice.[3] With an experimental setting resembling the real world, our paper is in line with the emerging framed field experiment literature (Ilomaki, 2012; Boulu-Reshef et al., 2016). Finally, our paper is original in two other aspects. We provide empirical evidence on judicial decision-making in a civil law institutional background, and we study the impact of child support guidelines. This contrasts with the previous literature which has mainly focused on the American case and on sentencing disparity in criminal cases.

Our main results are the following: the implementation of guidelines contributes to reducing the disparity between judges (i.e., the variance in similar cases is lower when the subjects have the opportunity to use the guidelines) but this effect is not systematic, with an increase in heterogeneity being observed in some specific cases.

[3] In 2010 the French Ministry of Justice produced a guideline for the determination of child support awards. This guideline is advisory; judges are free not to set the amount prescribed by the guideline.

The rest of the paper proceeds as follows. Section 1 provides an overview of prior works on the impact of guidelines on the elimination of inter-judge sentencing disparity. In Section 2, we describe the experiment and the methodology. In Section 3, we report our results. Section 4 concludes.

1. Literature review

The issue of inter-judge sentencing disparity has been widely discussed and researched over the past fifty years, by scholars in criminology, political science, law and economics. Comparatively little research has specifically examined the degree to which sentencing guidelines reduce sentencing disparity. Three waves of studies can be identified.

The first studies achieved somewhat mixed results. Following the introduction in the mid-1980s of mandatory guidelines in the United States, several studies sought to measure whether this reform reduced disparity of decisions between judges.[4] These studies revealed that overall sentencing disparity attributable to the judge either declined in the post-guidelines era or, conversely, increased in some district courts. For example, Anderson et al. (1999) used a definition of inter-judge disparity that measured the difference in the mean prison sentence for each judge relative to the mean prison sentence for all judges in the district. They found that sentencing disparity attributable to the judge declined substantially from 1986–1987 to 1988–1989 and remained relatively stable from 1990 to 1993. The expected difference in the lengths of sentences from two judges with comparable caseloads was 16–18 percent in the pre-guidelines era but only 8–13 percent in the post-guidelines era. The authors found that "the Guidelines have reduced the net variation in sentence attributable to the happen stance of the identity of the sentencing judge" (Anderson et al., 1999:303). The work of Hofer et al. (1999) came to the same type of conclusion. Using several techniques to determine whether inter-judge disparity had declined in the post-guidelines period, these authors concluded that the guidelines had had "modest success at reducing inter-judge disparity" for some types of offenses and in some of the districts examined (Hofer et al. 1999: 290–1). These results observed in criminal fields are confirmed in civil fields. Using data from the National Longitudinal Survey of Labor Market Experience of Youth, Argys et al. (2001) compared the variation in child-support awards for divorced or separated mothers living in states after guidelines were adopted to those living in states prior to the adoption of guidelines, controlling for sets of variables likely to account for differences. They showed that the introduction of guidelines had significantly reduced the disparity of amounts. However, they also showed that "the adoption of guidelines reduces the

[4] In criminal areas, the guidelines consist in proposing sentencing ranges by category of offense, according to the characteristics of the case (i.e., past criminal history of the offender and severity of the current offense). Judges are allowed to depart from this range (downward or upward) only under certain circumstances and provided that they give relevant reasons, which may relate to the characteristics of the case or legal issues (Schanzenbach et al., 2007).

likelihood of extreme amounts in some cases, but does not appear to improve horizontal equity in awards for the entire distribution of families" (Argys et al., 2001:246).

In contrast, other studies conclude that the variance attributable to the judge to whom the case was assigned increased in the post-guidelines period (Waldfogel, 1998; Payne, 1997; Lacasse and Payne, 1999). For instance, in the context of guilty pleas in the United States, Lacasse and Payne (1999) examined whether the variability of sanctions (prison sentences) attributable to judges had been eliminated by the use of guidelines. Their study shows that, once selection biases related to the characteristics of the judges and defendants were addressed, this variability would have increased. Specifically, the amount of variation attributable to the judge in trial decisions was 4% before reform and between 5 and 13% after reform, depending on the type of crime and the court. These authors concluded that the guidelines were probably insufficient to eliminate any variability in sanctions between judges, since judges retain a margin of discretion even in the presence of mandatory guidelines. In this study, the authors believe that judges may indeed be sensitive to the accused's remorse or may assess the evidence differently.

Other works have shown that the implementation of guidelines does not prevent extra-legal factors from influencing judicial sentencing, resulting in a continuance of inter-judge disparity. The sentencing guidelines prescribe that sentences are to be set by the judge within a range determined in accordance with a computed criminal history score and a computed offense score, and judges are explicitly forbidden from considering factors such as race, gender, socioeconomic status or family circumstances. However, empirical findings reveal that sentencing disparities according to race, gender, education, and socioeconomic status are prevalent in the federal criminal justice system. Abrahms et al. (2012) find evidence of significant inter-judge disparity in the racial gap in incarceration rates, thus supporting the model whereby at least some judges treat defendants differently on the basis of their race. Sorensen et al. (2012) show that personal circumstances do in fact figure in the determination of sentences and that racial and gender-based sentencing disparities remain even after accounting for personal circumstances, the criminal history score and the severity of current offense score: judges punish white women less severely and black men more severely than white men.

Other studies have looked at the effect on inter-judge disparity of relaxing the mandatory nature of the guidelines, further to the Supreme Court's decision in Booker v. United States in 2005 which greatly increased the degree of judicial discretion. Empirical work on the impact of Booker suggests that there have been increases in inter-judge sentencing disparities (Scott, 2010; Yang, 2015). Additionally, researchers have found that the increased judicial discretion since Booker has led to large and robust increases in racial disparities (Fischman et al., 2012; Yang, 2015). Thus the

loosening of the binding nature of guidelines in criminal matters in the United States has led to an increase in racial disparities in sentencing. Several authors (Fischman et al., 2012; Yang, 2015; Rehavi and Starr, 2014) point out that the increase in racial disparities in sentencing observed since Booker is undoubtedly the result of greater discretion for judges but also the consequence of the prosecutorial choices that preceded the judicial sentencing decisions. Their findings suggest that prosecutors would respond to increased judicial discretion after Booker by seeking binding mandatory minimum sentences for black defendants.

Taken together, the research on inter-judge disparity in US state and federal courts suggest that significant and nontrivial variations in sentences across judges remain even in jurisdictions with mandatory guidelines. These results suggest that constraining judicial discretion has not eliminated this inter-judge disparity.

2. Methods

We describe the design of the experiment in a first subsection (2.1.). Then we present our implementation and the sample (2.2.).

2.1. Design of the experiment

Following the typology proposed by Harrison and List (2004), our controlled experiment corresponds to a framed field experiment characterized by the selection of a panel of particular subjects (students at the ENM) and a significant contextualization of the task to be carried out (the setting of child support, as if they were judges).

The experimental protocol consisted in placing subjects in a situation quite similar to one they would encounter if they had to deal with divorce cases and focusing on the decision setting the amount of child support. We gave the subjects 48 vignettes, each one aiming at representing a frequent divorce situation involving children. These 48 situations were chosen on the basis of statistics from a representative database of divorce decisions in the first instance (CEEE-TGI, 2012).[5] We retained six criteria to characterize each of the 48 vignettes. Four correspond to the criteria of the national child support guidelines: the number of siblings for whom a child support award has to be set, the type of accommodation of the child, the income of the creditor parent,

[5] The database, compiled by the Ministry of Justice, contains detailed information on divorce decisions involving one or more children and parental separations with minor child(ren) taken in the first instance (*Tribunal de Grande Instance*, TGI), in June 2012. The database consists of 3,895 cases, involving 6,347 children. It is made representative at the national level through a weighting scheme based on the national statistics of the Ministry of Justice.

and the income of the debtor. The two other criteria are the child support proposal made by the creditor parent and the proposal of the debtor parent.

Two numbers of siblings were selected: one child and two children, representing 90% of divorce cases with children in the first instance.[6] Two accommodation situations were selected: main accommodation with the mother (every second weekend and half of the holidays with the father) and almost exclusive accommodation with the mother (the child has very little accommodation with the father). These two types of accommodation are decided in just over 7 out of 10 first-instance decisions, for a single child or two siblings.

For parental incomes, we searched the CEEE-TGI 2012 database for the most statistically typical "father-mother" income couples. The first combination is where both parents have relatively close intermediate incomes (between €1,200 and €2,000/month) but the father has a slightly higher income than the mother. For parents with one or two children living mainly with the mother, the median incomes of this type of combination are respectively around 1,600 euros per month for the father and 1,500 euros for the mother. The second combination consists in a relatively low female income although higher than the French minimum income,[7] and a much higher male income. The analysis of the cross-distribution of parental incomes led us to retain the following amounts: 1,000 euros per month for the mother and 1,900 euros per month for the father. In contrast, we added a third, rarer combination, in order to take into account situations where the mother earns more than the father. Using the same methodology as for the other cases, we selected a father's income equal to 1,100 euros and a mother's income equal to 2,500 euros per month.

Concerning the proposals by the parties,[8] the most frequent combinations are those with intermediate and fairly close amounts. Taking the median values, they result in an offer of 140 euros made by the father and a request of 200 euros made by the mother.[9] The second case is less frequent but contrasts with the previous case because of a more pronounced disagreement between the parents. The precise amounts retained correspond to the median values calculated from the cases corresponding to this type of combination of proposals, i.e., an offer equal to 100 euros and a request equal to 300 euros. Thirdly, we also retained the fairly frequent case where the father does not want to pay child support, so his offer is 0, and the mother asks for a relatively standard amount of child support, i.e. a median request equal to 150 euros. Finally, we chose to retain the fairly

[6] Whether or not the children are over the age of adulthood can affect the amount of support. In order to avoid this dimension being a source of heterogeneity, we specified the ages of the children using median ages according to the number of siblings (5 years for one child, 6 and 10 years for two siblings).
[7] In France, the minimum income (*Revenu de Solidarité Active*) is about €500/month for a single person.
[8] These proposals are crucial to judicial decision-making since, in French law, the judge cannot rule on what has not been requested by the parties (*extra petita*) and in particular cannot set an amount above the request of the creditor (*ultra petita*).
[9] Cases where the parents agree on the child support award are not considered in the experiment.

frequent situation where the father does not make a precise offer (he is willing to pay child support, but less than the mother's request, without explicitly indicating the precise amount) and the mother asks for a relatively standard amount of support, i.e., an undefined offer and a request equal to the median value of standard demands, i.e., 150 euros. In total we therefore retained 2 x 2 x 3 x 4 = 48 cases.

2.2. Implementation and sample

The experiment took place on October 13, 2017 with the 312 first-year students of the Ecole Nationale de la Magistrature (ENM). Each of the students was asked to decide on a child support amount for 48 different vignettes. Half of them (the treated group) had at its disposal the child support guidelines of the Ministry of Justice (in a simplified version, for quicker reading). On the contrary, the other (the control group) did not. None of the sub-group was aware that the experiment concerned the use of guidelines. A first session took place without guidelines, then a second session, organized without delay (to avoid contact between the students in the two sub-groups), took place with guidelines. The second sub-group was therefore unaware that the first had not benefited from guidelines. The distribution of the subjects in the two groups was based on pre-existing training groups set up by the pedagogical team at ENM, according to a logic of a background mix. This led to two sub-groups that were similar in terms of socio-demographic characteristics (see Table 1, columns 1 and 3). However, we applied a set of weights to further reduce small structural differences (see Table 1, columns 1 and 2).

Each of the 48 vignettes was summarized in a document which was very visual for ease of use, presenting the vignettes and including a box to enter the amount of child support. To test whether the order of cases could influence decisions, we produced four sets of answer sheets, each organized in a different order, and these sheets were distributed randomly to the students (see Appendix 1). The first group was asked to set an amount of child support without any comments on the existence of guidelines. The second group was asked to do the same exercise and received a simplified version of the Ministry of Justice's advisory guidelines. They were simply told that they were free to use them or not. In both cases, the introduction to the exercise was brief: a simple explanation of the context and how to fill in the document. We also explained that since the cases were very simplified, they had to consider that any information not given in the vignette, but which they could think of, should be considered identical from one case to another. No oral questions were accepted during the proceedings. No chatting with a neighbor was allowed. A few students asked clarification questions individually at the beginning of the exercise; we only answered them when they were questions of understanding and abstained when the question more or less

amounted to asking for help on the best way to proceed. The students completed the exercise with varying degrees of speed, but none of them ran out of time. Additionally, the students had to fill out a questionnaire on their personal background (age, gender, marital situation, curriculum, etc.) and preferences (altruism, risk, inequality[10]).

The 312 experiment sheets (one per student) had been filled in very correctly and practically without any missing data. The few missing data related to the information sheet and not to child support decisions. These missing data led us to discard two sets of responses (one in each of the two groups) and to make only 11 imputations for missing data (out of a total of 3,100 data items: 10 items on the information sheet * 310 students).

Table 1: Characteristics of the subjects, by sub-group

	Subjects with "guidelines" Non-weighted (1)	**Subjects with "guidelines" Weighted (2)**	**Subjects with "no guidelines" (3)**
Men	23.3%	27.5%	27.5%
Age	29.2	28.6	28.6
Married couple	13.3%	12.5%	12.5%
Couple with civil partnership contract	16.7%	10.6%	10.6%
In a common-law relationship	19.3%	25.6%	25.6%
Not in a couple	50.7%	51.3%	51.3%
With children	16.7%	16.3%	16.3%
Curriculum in law only	68.7%	68.1%	68.1%
Curriculum in law + other training	22.0%	25.0%	25.0%
No Curriculum in law	9.3%	6.9%	6.9%
Had been working prior to entering the ENM	42.7%	33.8%	33.8%
Had previously handled a divorce case	46.0%	50.0%	50.0%
Taste for selfishness	4.79	4.77	4.91
Taste for risk	4.17	4.10	4.04
Taste for equality	7.65	7.71	7.90
N	**151**	**151**	**161**

Source: ENM Guidelines Experiment database (2017).

[10] The selfishness scale proposed values between 0 and 10, with 10 corresponding to a purely selfish individual. The risk scale proposed values between 0 and 10, with 10 corresponding to a risk lover. The inequality aversion scale proposed values between 0 and 10, with 10 corresponding to an individual who is very much in favor of reducing inequalities. In the subsequent analyses, we used dummy variables which, for each scale, opposed two sub-groups of students: altruistic people (selfishness scale less than 5) *versus* selfish people; risk lovers (risk scale greater than 4) *versus* those who are risk averse; equality lovers (equality scale greater than 7) *versus* those who prefer inequality.

3. Results

We present the impact of the child support guidelines on the level of child support awarded (4.1.), on inter-judge disparity (4.2.) and with regard to the *ultra petita* procedural rule (4.3.).

3.1. The impact of the guidelines on the level of child support awarded

When child support guidelines were generalized across the United States, this policy was explicitly part of the fight against child poverty. It was well known that child support payments were on average low and that the expected impact of the guidelines was an increase in child support payments. In France, the guidelines were introduced in a completely different context. They were not specifically intended to combat child poverty, but rather to provide judges with a decision-making tool to facilitate divorce proceedings and avoid the heterogeneity created by the coexistence of multiple unofficial *ad hoc* guidelines. In this context, it was therefore difficult to predict whether the French guidelines would have an upward or downward effect.

On the basis of all 48 cases in the experiment, we observe that the average child support decided upon by the subgroup "with guidelines" was slightly higher than the average of the decisions made by the subgroup "without guidelines": 150.30 euros *versus* 146.60 euros, and this difference is statistically significant at the 5% level. This increase can be explained in part by the fact that most often the students spontaneously, i.e., when they were not aware of the guidelines, set amounts lower than the value suggested by the guidelines, resulting in an average increase in the amounts once the value suggested by the guidelines was known. This is in fact what we observed in our experiment in 26 cases out of 48, for which the average of the amounts fixed without (the possibility to use) the guidelines was lower than both the average of the amounts fixed with (the possibility to use) the guidelines and the amount suggested by the guidelines (in 20 cases out of 26 the difference between the two means is significant at the 5% level). Conversely, in only 20 cases the average of the amounts set without the guidelines was higher than both the average of the amounts set with the guidelines and the amount suggested by the guidelines (in 14 cases the difference between the two means is significant at the 5% level).[11] Therefore, it is questionable whether students in the treatment group would have been significantly influenced by the guidelines only when dealing with a case with specific characteristics. The econometric analysis presented in Table 2 helps to answer this question.[12]

[11] The last two cases correspond to situations where the average of the amounts set without the guidelines is higher than the value suggested by the guidelines, but lower than the average of the amounts set with the guidelines, whereas one would expect it to be higher (but the difference between the two means is not significant at the 5% level).

[12] Since each of the 48 responses was processed by the same student (and this for each of the 310 students), our estimates may suffer from a non-independence bias and in this case the standard deviations of the estimators would be underestimated, giving the

Table 2. Econometric estimation of the amount of child support

Constant	154.38 ***
Student characteristics	
With guidelines	0.16
Without guidelines	Ref.
Woman	0.24
Male	Ref.
Age	-0.12
In a couple, married or not	-0.45
Not in a couple	Ref.
With child(ren)	0.49
No children	Ref.
Only law school	4.58 ***
No law school background or law school + other education	Ref.
Worked before ENM	-1.28
Did not work before ENM	Ref.
Previously handled a divorce case	-0.36
Never handled a divorce case	Ref.
Altruist	-0.49
Selfish	Ref.
Risk lover	1.10 *
Risk-averse	Ref.
Equality lover	2.43 ***
Equality-averse	Ref.
Lot 1	10.07 ***
Lot 2	18.69 ***
Lot 4	13.44 ***
Lot 3	Ref.
Characteristics of the 48 vignettes	
Almost exclusive accommodation with the mother	17.39 ***
Main accommodation with the mother	Ref.
Siblings of two children	-13.31 ***
One child	Ref.
Proposals "unspecified offer – 150"	-31.93 ***
Proposals "0 – 150"	-33.91 ***
Proposals "100 – 300"	2.56 *
Proposals "140 – 200"	Ref.
Incomes "1,100- 2,500"	-40.53 ***
Incomes "1,900 – 1,000"	24.15 ***
Incomes "1,600 – 1,500"	Ref.
Interactions between Guidelines and vignette characteristics	
Guidelines * Almost exclusive accommodation with the mother	5.67 ***
Guidelines * Siblings of two children	-2.11 *
Guidelines * Proposals "unspecified offer –150"	5.30 ***
Guidelines * Proposals "0 – 150"	8.10 ***
Guidelines * Proposals "100 – 300"	-3.05 *
Guidelines * Incomes "1,100 – 2,500"	-10.87 ***
Guidelines * Incomes "1,900 – 1,000"	9.13 ***
N	14,876
R^2	58.9%

Source: ENM Guidelines Experiment database (2017). ***: significant at the 0.1% threshold. **: significant at the 1% threshold. *Significant at the 5% threshold. Estimations made with an OLS regression. Average amount of child support: €149.

illusion that such effect is significant when it is not. That is why we tested a multilevel model where each of the 310 sets of 48 responses was "nested" within the 310 students. The results obtained (available from the authors) are very close to those obtained with the ordinary least squares regression presented in Table 2. It should also be noted that the individual characteristics of students had no substantial influence on decision-making behavior, with the exception of their educational background: those who only attended law school were significantly more generous. And students who had a greater preference for inequality reduction than other students were also somewhat more generous (see Table 2).

On average, and all things being equal, when dealing with a case of almost exclusive accommodation with the mother, students in the "control group" set an amount about 17 euros higher than in cases of main accommodation with the mother, i.e., one ninth of the mean child support. However, in the same situation, students "with guidelines" increased the child support even more (23 euros), and the difference (attributable to the possibility of using the guideline) of 6 euros is highly statistically significant ($p < 0.0001$) but substantially very low: it amounts to less than 4% of the mean child support.

With regard to the number of siblings, the two sub-groups made their decision in a more similar way: in almost the same proportion (the estimated difference is small – 2 euros – and is only significant at the 5% threshold), they awarded lower child support when they dealt with a case with two children than with a case with one child. The use of the guidelines would therefore not have a major impact with respect to number of siblings.

Compared to the fairly common situation where the parties made intermediate and not very different proposals (140-200), the students, irrespective of the sub-group, set significantly lower amounts when the request was low (150 euros). This undercutting behavior was influenced by the possibility of using the guideline, since "treated" students undercut significantly less (about 5 and 8 euros respectively for the two cases with an offer equal to 150) than "control" students. With an "unspecified offer-150" pair of proposals, the reduction for the sub-group of "control" students is estimated at -32 euros and for the sub-group of students "with guidelines" at -27 euros (p-value < 0,1%). With a "0-150" pair of proposals, the reduction for the sub-group of "control" students is estimated at -34 euros and for the sub-group of students "with guidelines" at -26 euros (p-value < 0,1%). However, compared to the cases with a "140-200 pair of proposals", this reduction behavior was not observed when the students dealt with cases with proposals equal to "100-300".

Finally, as for the parents' income, two opposite behaviors were observed. Compared to the situation where both parents had intermediate and similar incomes (1,600-1,500), the students set higher child support amounts in cases where the father earned significantly more than the mother (1,900-1,000)[13] and lower amounts in the opposite case (1,100-2,500).[14] But what is interesting to note is that these two behaviors were reinforced by the possibility of using the guidelines, with the differences (+9 euros and -11 euros) between the two subgroups being highly statistically significant, while not substantially very large: these differences amount, respectively, to only 6% and 7% of the mean child support.

[13] +24 euros for the "control" sub-group and +33 euros for the "treated" sub-group.
[14] -41 euros for the "control" sub-group and -51 euros for the "treated" sub-group.

3.2. The impact of the guidelines on the disparity of child support amounts

The variance calculated across all child support amounts was greater when students could use the guidelines than when they could not (2,498 *versus* 2,272), but this positive difference in variance was due solely to the difference in variance between cases (between vignettes).[15] Conversely, the variance within cases was lower when students had the possibility of using the guidelines than when they did not (817 *versus* 1,048). This global negative difference seems to validate the hypothesis that the use of guidelines would reduce the disparity of legal decisions on child support amounts. However, it should be pointed out that a lower variance with guidelines was not observed for all the cases, but only in 35 out of 48 cases.[16] This non-systematism (similar to what we observed with regard to the impact on the level of child support amount, see above) led us to investigate why, for certain types of cases, the use of the guidelines generated an increase in the heterogeneity of decisions. To do this, we estimated the differences "with _ without guidelines" in variance of the child support decision as a function of the characteristics of the cases (Table 3)[17].

The first lesson that can be drawn from this regression (Table 3, column 1) is that the impact of the guidelines would not differ according to the criteria relating to children (number of siblings and type of accommodation). With regard to the parties' proposals, it can be observed that the difference in effect (compared to the reference situation "140-200") of the potential use of the guidelines on the variance is only significant for the "100-300" pair of proposals. Thus, in the event of very different proposals, the potential use of the guidelines significantly reduces the variance more than it does for proposals that are close to each other: the variance is reduced by 576 more than in the reference, a value equal to 25% of the average variance observed when students did not have the opportunity to use the guidelines.[18]

---

[15] Using statistical indicators of influence, we show that the magnitude of these variances is not due to a few students who would have made decisions quite systematically and very differently from those of other students (outliers). And the individual characteristics of the students would not be statistically related, *ceteris paribus*, to their individual contribution to the variance. The results of these additional analyses are available from the authors.

[16] The negative difference is significant at the 5% threshold for 28 out of 35 cases and the positive difference is significant for 9 out of 13 cases. Thus, in 11 cases out of 48 there is no statistically significant difference.

[17] Variance is a metric-dependent indicator of heterogeneity, so it is expected that the variance of high child support decisions will be mechanically higher than the variance of lower child support decisions. A case characteristic (e.g., the father's income) can therefore be related to the amount of support, and therefore to the variance of decisions in that case and ultimately to the difference in "with *versus* without guidelines" variances. If a significant regression coefficient is found between a case characteristic and the difference in variance, we do not know whether it is an impact on this difference or on the amount of child support. To avoid this drawback, one can use the coefficient of variation (a unitless measure of dispersion: the variance is divided by the variable mean), which measures relative heterogeneity, rather than variance. We therefore duplicated all the estimates presented in the article using this other indicator of dispersion (results available from the authors). The results are very close to those obtained with variance, which justifies not mentioning them in the text.

[18] Further calculations allow us to note that the difference (in the effect of the guidelines on the variance) is also statistically different (at the 1% threshold) between this pair of proposals (100-300) and respectively the other two pairs of proposals, "0-150" and "unspecified offer-150". On the other hand, there would be no significant difference between these last two pairs of proposals.

Table 3: estimations of the difference "with – without guideline" of variance of child support decisions

	(1)	(2)
Constant	-327	-371
Almost exclusive accommodation with mother	295	295*
Main accommodation with the mother	Ref.	Ref.
Siblings of two children	-132	-132
One child	Ref.	Ref.
Proposals "unspecified offer –150"	110	/
Proposals "0 – 150"	81	/
Proposals "100 – 300"	-576*	/
Proposals "140 – 200"	Ref.	/
Incomes "1,100 – 2,500"	-278	/
Incomes "1,900 – 1,000"	656**	/
Incomes "1,600 – 1,500"	Ref.	/
Incomes "1,600 – 1,500" + Proposals "100 – 300"	/	-512
Incomes "1,600 – 1,500" + Proposals "0 – 150"	/	118
Incomes "1,600 – 1,500 » + Proposals "unspecified offer – 150"	/	185
Incomes "1,900 – 1,000 » + Proposals "100 – 300"	/	-559
Incomes "1,900 – 1,000 » + Proposals "0 – 150"	/	1,202***
Incomes "1,900 – 1,000 » + Proposals "unspecified offer – 150"	/	1,228***
Incomes "1,900 – 1,000 » + Proposals "140 – 200"	/	543
Incomes "1,100 – 2,500 » + Proposals "100 – 300"	/	-147
Incomes "1,100 – 2,500 » + Proposals "0 – 150"	/	-569
Incomes "1,100 – 2,500 » + Proposals "unspecified offer – 150"	/	-573
Incomes "1,100 – 2,500 » + Proposals "140 – 200"	/	-34
Incomes "1,600 – 1,500 » + Proposals "140 – 200"	/	Ref.
N	48	48
R^2	50.0%	77,4%

Source: ENM Guidelines Experiment database (2017). ***: significant at the 0.1% threshold. **: significant at the 1% threshold. *: Significant at the 5% threshold. Regression by ordinary least squares. Average difference of variance of child support decisions: -216.

When the parents' incomes were highly unequal and in favor of the father (1,900-1,000), the possibility of using the guidelines led to an increase in variance which is very significantly different from the reduction in variance observed for the couples with similar incomes (1,600-1,500): +656, i.e., a value equal to 29% of the average variance observed in the situation where students did not have the opportunity of using the guidelines.[19]

3.3. Heterogeneity of decisions and procedural rule

In French law, there is a procedural rule according to which a judge must rule on everything that is requested and only on what is requested. In practice, this results in the requirement to make a decision within the range of proposals (except in very special situations which must be justified by the judge in her decision). In the case of child support, this means that the judge must choose an amount between the offer and the request expressed by the parties. To go further in explaining the

[19] Additional calculations show that the difference is also significant between the "1,100-2,500" and the "1,900-1,000" income couples.

variability of the effect of the guidelines on the heterogeneity of decisions, it may be interesting to study the impact of this procedural rule. Indeed, when the guidelines suggest an amount outside the range of proposals, it may be thought that this encourages some judges, but not all, to depart from this rule. It may therefore be assumed that a suggestion from the guidelines outside the parties' proposals is a source of heterogeneity.

We explored this idea using the regression presented in Table 3, column 2, which specifies all possible combinations of income and proposal pairs, including those where the value suggested by the guidelines is either higher than the child support requested by the mother (which is low with respect to income) or lower than the child support proposed by the father (which is high with respect to income).[20]

The suggestion from the guidelines was systematically (regardless of the number of children and type of accommodation) greater than the request when the couple had the 1,900-1,000 income type and the request was equal to 150.[21] In such a configuration, our hypothesis therefore leads us to expect an effect of increasing the variance associated with the possible use of the guideline. This effect is indeed estimated as positive and very significant (1,002 and 1,228; p-value < 0,1%). The magnitude of these two effects should be compared to the average variance observed when the students did not have the opportunity to use the guidelines, i.e., at 2272: the magnitude is more than half. Conversely, the suggestion from the guidelines was almost systematically lower than the offer when the couple had the 1,100-2,500 income pair and the offer was equal to 140 or 100.[22] In this case, an increase in the variance associated with the use of the guidelines was again expected to be observed, but this positive effect is not confirmed by the estimate.

In our experiment, *ultra petita* decisions were quite rare when the students were dealing with cases without guidelines (3%). They were more frequent when they had the opportunity to use the guidelines (9%). If we focus on the 14 vignettes where the value of child support suggested by the guidelines was higher than the request and therefore may have encouraged an *ultra petita* judgement, what did we observe? According to the results from the estimation of the probability of *ultra petita* decisions (Appendix 2), it seems that the incentive given by the guidelines to disregard the procedural rule acted on all individuals rather than just on those with particular characteristics. We can see that the probability of deviating from the procedural rule is positively related to the risk aversion indicator. Such a result may be interpreted by the fact that risk-loving individuals are ready

[20] The experimental design consists of 14 cases where the amount suggested by the guidelines is greater than the mother's request, 7 cases where the amount suggested by the guidelines is less than the father's offer, and 27 cases where the value suggested by the guidelines is between the two parental proposals.
[21] Depending on the other characteristics of the cases, the values suggested by the guidelines are €164, €192, €211 or €257.
[22] Depending on the other characteristics of the cases, the values suggested by the guidelines are €72, €84, €97 or €113.

to disregard the procedural rule because they do not fear that their decision will be overturned by a higher authority, such as an appeal court in our case. Nevertheless, the significance level of that effect is quite low (5%), which leads us to favor the hypothesis of an indistinct incentive effect of the guideline.

4. Conclusion

This research presents a framed field experiment on the effect of child support guidelines on inter-judge disparities. This experiment produced several salient results. Firstly, we show that the impacts of an advisory guideline are of different magnitudes and of different signs (the variance with guidelines may be higher or lower than without guidelines). Next, on average (i.e. for the 48 typical cases considered simultaneously), we find evidence that advisory guidelines reduce inter-judge disparities, since we show that the intra-group variance is lower when the subjects have the opportunity to use the guidelines. Nevertheless, we observe that this effect is not systematic, since we observe increases in inter-judge disparities in some cases. These cases are characterized by the presence of a significant income gap within the couple to the benefit of the debtor and a modest request expressed by the creditor. We interpret this result as being due to some decision-makers correcting the apparent inconsistency between a debtor's high income and a low child support request comparatively to what the guidelines suggest, while others validate the amount claimed by the creditor on the basis that a judge cannot judge *ultra petita*.

These results present two main limitations. Firstly, the subjects in the experiment were not judges, but ENM students. As a result, it is likely that some of the observed effects were exaggerated, comparatively to what could be observed in a real judicial context. In particular, a significant proportion of subjects ruled outside the range of the parties' proposals, while this situation is extremely rare in judicial decisions. Nevertheless, even if this type of behavior was over-represented, it gives us some clues as to why the introduction of guidelines can lead to an increase in inter-judge variation. Second, for organizational reasons, we opted for an experiment with two sub-groups of subjects, one ruling without guidelines and the other with. This option constitutes a limitation because it restricted our statistical exploitations by forcing us to do analyses of differences in average subgroup decisions. It would have been more relevant to analyze differences in individual decisions. This would have required all the subjects to rule successively without and then with guidelines on the forty-eight typical cases, which was not technically possible.

References

Abrams, D. S., Bertrand M., Mullainathan S. (2012), "Do Judges Vary in Their Treatment of Race?" *Journal of Legal Studies*, 41, 347-383.

Anderson J. T., Kling J. R., Stith K. (1999), "Measuring Inter-judge Sentencing Disparity: Before and After the Federal Sentencing Guidelines", *Journal of Law and Economics*, 42(1), 271-307.

Argys L. M., Peters H. E., Waldman D. M., (2001), "Can the Family Support Act Put Some Life Back into Deadbeat Dads? An Analysis of Child-Support Guidelines, Award Rates, and Levels", *The Journal of Human Resources*, 36(2), 226-252.

Boulu-Reshef B., Comeig I., Donze R., Weiss G. D. (2016), "Risk aversion in prediction markets: A framed-field experiment", *Journal of Business Researc*h, 69(11), 5071-5075.

Cohen A., Yang C. S. (2019), "Judicial politics and sentencing decisions", *American Economic Journal: Economic Policy*, 11, 160–91.

Fischman, J. B., Schanzenbach M. M. (2012), "Racial Disparities under the Federal Sentencing Guidelines: The Role of Judicial Discretion and Mandatory Minimums", *Journal of Empirical Legal Studies*, 9, 729–64.

Guthrie C., Rachlinski J. J., Wistrich A. J. (2001), "Inside judicial mind", *Cornell Law Review*, 86, 778-830.

Harrison G. W., List. J. A. (2004), "Field Experiments." *Journal of Economic Literature*, 42 (4), 1009-1055.

Hofer P. J., Blackwell K. R., Ruback R. B. (1999), "The effect of Sentencing Guidelines on Inter-Judge Sentencing Disparity", *The Journal of Criminal Law and Criminology*, 90(1), 239-321.

Ichino A., Polo M., Rettore E. (2003), "Are Judges Biased by Labor Market Conditions?" *European Economic Review*, 47(5), 913-944.

Ilomäki I. (2012), "Framed Field Experiment with Stock Market Professionals, *Journal of Behavioral Finance*, 13(4), 251–258.

Kahan D. M., (2015), "Laws of Cognition and the Cognition of Law", *Cognition*, 135, 56-60.

Lacasse C., Payne A. A., (1999), "Federal Sentencing Guidelines and Mandatory Minimum Sentences: Do Defendants Bargain the Shadow of the Judge?", *The Journal of Law & Economics*, 42(S1), 245-270.

Liu Z., (2018), "Does Reason Writing Reduce Decision Bias? Experimental Evidence from Judges in China", *Journal of Legal Studies*, 47, 83-118.

Marinescu I. (2011), "Are judges sensitive to economic conditions? Evidence from UK employment tribunals", *Industrial and Labor Relations Review*, 64(4), 673–698.

Payne A. (1997), "Does Inter-Judge Disparity Really Matter? An Analysis of the Effects of Sentencing Reforms in Three Federal District Courts", *International Review of Law and Economics*, 17, 337-366.

Peresie J. L. (2005), "Female Judges Matter: Gender and Collegial Decision making in the Federal Appellate Courts", *The Yale Law Journal*, 114, 1759-1790.

Rehavi M. M., Starr S. B., (2014), "Racial Disparity in Federal Criminal Sentences", *Journal of Political Economy*, 122(6), 1320-1354.

Schanzenbach M. M., Tiller E. H. (2007), "Strategic Judging under the U.S. Sentencing Guidelines: Positive Political Theory and Evidence." *Journal of Law, Economics, and Organization,* 23(1), 24–56.

Scott R. W. (2010), "Inter-judge sentencing disparity after Booker: A first look", *Stanford Law Review,* 63, 1–66.

Sorensen T., Sarnikar S., Oaxaca R. L., (2012), "Race and Gender Differences Under Federal Sentencing Guidelines", *The American Economic Review*, 102(3), 256-260.

Spamann H., Klöhn L. (2016), "Justice Is Less Blind, and Less Legalistic, Than We thought: Evidence from an Experiment with Real Judges", *Journal of Legal Studies*, 45, 255-280.

Waldfogel J. (1998), "Does Inter-Judge Disparity Justify Empirically Based Sentencing Guidelines?", *International Review of Law and Economics*, 18, 293–304.

Wistrich A. J., Rachlinski J. J., Guthrie C. (2015), "Heart Versus Head: Do Judges Follow the Law or Follow Their Feelings", *Texas Law Review*, 93, 855-923.

Wooldredge J. (2010), "Judges' Unequal Contribution to Extralegal Disparities in Imprisonment", *Criminology*, 48(2), 539-567.

Yang C. S. (2015), "Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing", *The Journal of Legal Studies*, 44(1), pp. 75-111.

Appendix 1 -The 48 vignettes given to the student magistrate subjects

		1 child, 5 years old		2 children, 6 and 10 years old	
Income	Proposals	Main accommodation with the mother	Almost exclusive accommodation with the mother	Main accommodation with the mother	Almost exclusive accommodation with the mother
Father: 1,900€ Mother: 1,000€	Father: 100 Mother: 300	**1**	**5**	**25**	**29**
	Father: 0 Mother: 150	**2**	**6**	**26**	**30**
	Father: unknown Mother: 150	**3**	**7**	**27**	**31**
	Father: 140 Mother: 200	**4**	**8**	**28**	**32**
Father: 1,100€ Mother: 2,500€	Father: 100 Mother: 300	**9**	**13**	**33**	**37**
	Father: 0 Mother: 150	**10**	**14**	**34**	**38**
	Father: unknown Mother: 150	**11**	**15**	**35**	**39**
	Father: 140 Mother: 200	**12**	**16**	**36**	**40**
Father: 1,600€ Mother: 1,500€	Father: 100 Mother: 300	**17**	**21**	**41**	**45**
	Father: 0 Mother: 150	**18**	**22**	**42**	**46**
	Father: unknown Mother: 150	**19**	**23**	**43**	**47**
	Father: 140 Mother: 200	**20**	**24**	**44**	**48**

Lot 1: 1, 2, 3, 4, 5, 6, 7, 8 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48.

Lot 2: 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 1, 2, 3, 4, 5, 6, 7, 8, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 25, 26, 27, 28, 29, 30, 31, 32.

Lot 3: 3, 2, 1, 4, 7, 6, 5, 8, 11, 10, 9, 12, 15, 14, 13, 16, 19, 18, 17, 20, 23, 22, 21, 24, 27, 26, 25, 28, 31, 30, 29, 32, 35, 34, 33, 36, 39, 38, 37, 40, 43, 42, 41, 44, 47, 46, 45, 48.

Lot 4: 11, 10, 9, 12, 15, 14, 13, 16, 19, 18, 17, 20, 23, 22, 21, 24, 3, 2, 1, 4, 7, 6, 5, 8, 35, 34, 33, 36, 39, 38, 37, 40, 43, 42, 41, 44, 47, 46, 45, 48, 27, 26, 25, 28, 31, 30, 29, 32.

Appendix 2. Estimation of the probability of deciding *ultra petita*

Constant	0.96
Student characteristics	
Age	-0.14
Woman	0.23
Male	Ref.
In a couple, married or not	-0.10
Not in a couple	Ref.
With child(ren)	-0.20
No children	Ref.
Only law school	0.22
No law school background or law school + other education	Ref.
Worked before ENM	1.12
Did not work before ENM	Ref.
Previously handled a divorce case	-0.80
Never handled a divorce case	Ref.
Altruist	-0.05
Selfish	Ref.
Equality lover	0.14
Equality-averse	Ref.
Risk lover	1.19 *
Risk-averse	Ref.
Lot 1	-2.07 *
Lot 2	0.61
Lot 4	1.97 *
Lot 3	Ref.
Vignettes (definitions on appendix 1)	
Case #2	1.31 **
Case #3	1.22 **
Case #6	2.44 ***
Case #7	2.10 ***
Case #8	0.57
Case #22	0.29
Case #23	-0.61
Case #26	1.31 **
Case #27	0.94 *
Case #30	2.52 ***
Case #31	2.35 ***
Case #32	0.19
Case #46	0.67
Case #47	Ref.
N	2,100

Source: ENM Guidelines Experiment database (2017). Sample: decisions made by students in the "treated group" who had to deal with cases where the value suggested by the guidelines was greater than the request. For a precise definition of the cases according to their number, see Appendix 1. The estimation of this binary logit is done with a multilevel model where random effects are assumed at the student level (ie the group level, nesting the cases).

. ***: significant at the 0.1% threshold. **: significant at the 1% threshold. *: significant at the 5% threshold.