Dhami, M. K. & Belton, I. (in press). On getting inside the judge's mind. *Translational Issues in Psychological Science.*

On Getting Inside the Judge's Mind

Mandeep K. Dhami[1] and Ian K. Belton[2]

[1]Professor of Decision Psychology, Middlesex University, UK

[2]PhD student, Middlesex University, UK

Send Correspondence to:
Ian K Belton
Middlesex University
The Burroughs
Hendon
London, NW4 4BT
UK
E-mail: IB271@live.mdx.ac.uk

Public significance statement

We propose that researchers studying judicial decision-making ought to examine decisions made on real(istic) cases using a representative experimental design, and they should analyze individual judge or bench decision data using psychologically plausible models. This will make the research more relevant to the judiciary and thus make it more difficult for judges and legal policy-makers to ignore the findings.

Abstract

According to the scales of justice, the judge, in an unbiased way and directed by law, attends to all of the available information in a case, weighs it according to its significance, and integrates it to make a decision. By contrast, research suggests that judicial decision-making departs from the cognitive balancing act depicted by the scales of justice. Nevertheless, the research is often dismissed as irrelevant, and the judiciary, legal policy-makers and the public remain largely unconvinced that the status quo needs improving. One potential rebuttal to the scientific findings is that they lack validity because researchers did not study judges making decisions on real cases. Another potential argument is that researchers have not pinpointed the psychological processes of any specific judge because they analyzed data over judges and/or used statistical models lacking in psychological plausibility. We review these two grounds for appeal against the scientific research on judicial decision-making, and note that it appears researchers' choices of data collection methods and analytic techniques may, indeed, be inappropriate for understanding the phenomena. We offer two remedies from the sphere of decision-making research: collecting data on judicial decision-making using representative design, and analyzing judicial decision

data using more psychologically plausible models. Used together, we believe these solutions can help researchers better understand and improve legal decision-making.

The scales of justice represent an ancient and near universal depiction of the judicial decision-making process. According to this metaphor, judges as triers of fact, in an unbiased way and directed by the law, carefully attend to all of the available information in a case, weigh it according to its significance for the issue at hand, and integrate it to make a decision. A judge's ability to perform this cognitive balancing act when making highly consequential decisions is almost accepted as a given: after all, judges are recruited on the basis of their "sound judgment" (e.g., see Judicial Appointments Commission, 2011). And, when judicial decisions are challenged, it is typically not on the basis of a judge's poor or biased decision-making but rather on the basis of some misapplication of law, procedural mistake or error in holding that the evidence supported the outcome (Cohen, 2006). Finally, judges themselves are highly confident in their decision-making (Dhami & Ayton, 2001).

For decades, however, there has been a constant trickle of scholarly research in several disciplines including psychology, law, and criminology, suggesting that judicial decision-making departs from the rational ideal depicted by the scales of justice. For instance, past research has found that judicial decisions are affected by a myriad of extra-legal factors such as those related to the personal characteristics of the defendant or plaintiff including his/her gender, race and age (e.g., Doerner & Demuth, 2010; Manning, Carroll, & Carp, 2004; Mitchell, 2005; Rachlinski, Johnson, Wistrich, & Guthrie, 2009; Robbennolt, 2002; Turner & Johnson, 2006). Judges' own gender, race and age have also been found to influence their decisions (e.g., Wooldredge, Griffin, & Thistlethwaite, 2013; Chew & Kelley, 2008; Coontz, 2000; Kulik, Perry, & Pepper, 2003; Martin & Pyle, 2004; Peresie, 2004). Studies have also demonstrated that judicial decisions are

distorted by cognitive illusions such as framing effects, anchoring, and hindsight bias (e.g., Englich, Mussweiler, & Strack, 2006; Guthrie, Rachlinski, & Wistrich, 2001; Hastie & Viscusi, 1998; Rachlinski, Guthrie, & Wistrich, 2011) and biased by non-cognitive factors such as hunger (Danziger, Levav, & Avnaim-Pesso, 2011). In addition, there is evidence of intra-judge inconsistency, i.e., a judge's decisions differ from test to re-test, and inter-judge inconsistency, i.e., judges disagree on a decision for the same case (e.g., Collins, 2008; Dhami & Ayton, 2001; Robbennolt, 2002). Finally, studies show that judges may use simple heuristic strategies to make their decisions that ignore much of the available, relevant information (e.g., Dhami, 2003; Dhami & Ayton, 2001; Guthrie et al., 2001; Englich et al., 2006; von Helversen & Rieskamp, 2009).

Although the scientific research suggests that the notion of judges performing a cognitive balancing act is a myth, the judicial community has seemingly managed to dismiss the empirical evidence as irrelevant. Legal policy-makers and the public also appear to remain largely unconvinced that judicial decision-making needs much improvement. Indeed, little reference is made to empirical research findings either in media articles criticizing judicial decisions, in judicial training, or during the development of legal policies.

One of the main rebuttals to the scientific findings is that the research lacks internal and external validity (i.e., arguments that researchers do not measure what they claim to be measuring, and that which exists beyond the laboratory) because researchers did not study judges making decisions on real cases. Here, the lack of realism or representativeness of the case stimuli presented to judges is of key concern.

Another potential argument against the scientific findings is that they lack relevance to any specific judge because researchers have analyzed data aggregated over groups of judges and/or used statistical models lacking in psychological plausibility to describe judges' cognitive

processes. This criticism highlights that judicial decision-making research is ultimately a psychological undertaking and so the psychological applicability and plausibility of models used to capture such decisions is important. In addition, being able to pinpoint individual judges' decision strategies means that research can be used to hold judges to account.

In the present paper, we review these two grounds for appeal against the scientific research on judicial decision-making. We demonstrate that researchers' traditional choice of data collection methods and analytic techniques may indeed be inappropriate for understanding judicial decision-making. We offer two alternatives for future research: representative design for collecting data on judicial decision-making, and more psychologically plausible models for analyzing judicial decision data. We conclude that when used together, this data collection method and analytic technique can help researchers better understand judicial decision-making.

Data Collection Methods and Analytic Techniques

Researchers investigating judicial decision-making have used a variety of data collection methods (see Dhami & Belton, 2015). These include interview and questionnaire surveys of judges who report their decision-making strategies (e.g., Harris & Jesilow, 2000); observations of court hearings (e.g., Dhami, 2003); document analyses of court records (e.g., Baumer, Messner, & Felson, 2006); analysis of official court statistics (e.g., Merrall, Dhami & Bird, 2010); and experiments with judges deciding on hypothetical cases (e.g., Dhami & Ayton, 2001).

Researchers have also used a variety of data analytic techniques beyond descriptive statistics such as tests of mean differences (e.g., Rachlinski, Guthrie, & Wistrich, 2011) and correlational and predictive (regression) statistics (e.g., Merrall et al., 2010). Next, we evaluate these data collection methods and analytic techniques in the study of judicial decision-making.

'Researchers Did Not Study Judges Making Decisions on Real Cases'

The argument that research findings lack validity because researchers did not study judges making decisions on real cases can be applied to interview and questionnaire surveys of judges and experiments on judges. Beyond the failures of memory, there are well-known problems arising from asking people to report their decision-making strategies. Among other things, data may be unreliable and invalid because of social desirability response bias (Paulhus, 1991). People also have difficulty introspecting about their cognitive processes (Nisbett & Wilson, 1977; but see Newell & Shanks, 2014). In addition, interview and questionnaire surveys of judges cannot answer causal questions about the factors that influence judicial decisions.

Although experimental studies can identify cause-effect relationships, they are often criticized for lacking validity. In 'systematic' experimental design, the researcher selects one or a few independent variables of interest (e.g., defendant's gender and offence) and manipulates them so they vary systematically while holding other extraneous and potentially confounding variables constant or allowing them to vary randomly. The researcher then measures the resulting changes in the dependent variable(s) (e.g., judge's sentencing decision). Here, one can imagine creating artificial cases or rare combinations of factors that judges are not experienced in dealing with (e.g., elderly, female defendants being sentenced for violent or sexual offences), and so studies may not elicit natural response patterns (i.e., lack internal validity; see e.g., Hammond & Stewart, 1974; Moore & Holbrook, 1990; Phelps & Shanteau, 1978). We address this issue further in our discussion of representative design.

The findings of research involving judges making decisions on hypothetical cases may also be difficult to generalize to cases beyond the laboratory situation (i.e., lacking in external validity). The hypothetical cases typically used in experimental studies are necessarily brief and lack the richness of detail present in real life cases. Concern over generalizability is particularly

great when researchers study the effect of only one variable, often resulting in judges making decisions on a single case. Later, in our discussion of representative design, we consider how researchers can examine cause-effect relations between information presented to a judge and his/her decisions, without threatening the internal and external validity of the research findings.

Clearly, the argument that research findings lack internal and external validity because researchers did not study judges making decisions on real cases does not apply to studies involving courtroom observations and analyses of court records and statistics. However, this does not imply that these are the most appropriate methods to use. This is because although these methods involve judges deciding on real cases, they are limited in their ability to determine the cause-effect relations between information presented to judges and their decisions. At most, these methods can provide evidence of the factors that are associated with judicial decisions. In addition, courtroom observational studies tend to be limited in their ability to reliably and validly observe all information presented to a judge and/or his/her decision behavior. Similarly, the findings of studies involving data recorded in court records or published in official statistics may be limited due to the lack of relevant data being available from these sources.

'Researchers Analyzed Data Aggregated Over Groups of Judges and/or Using Psychologically Implausible Statistical Models'

The argument that research findings lack relevance to any specific judge because researchers have analyzed data aggregated over groups of judges and/or used statistical models that lack psychological plausibility holds true. Analysis of judicial decision-making data using the nomothetic tradition is commonplace in studies using real cases (e.g., studies based on officially published court statistics) and hypothetical ones (e.g., experimental studies). Group differences in decisions made on multiple cases or one case are often analyzed using means and

standard deviations, followed by regression analyses or analysis of variance. While nomothetic research has a role to play in understanding wider trends across a particular jurisdiction, it is unclear to what extent such studies can capture the decision strategy of any individual judge. This creates difficulties when seeking to persuade judges of the relevance of a given study to their practice. In addition, it is likely that individual judges have different judgment strategies from one another, and may differ in the strategy they use from one case to another (e.g., Dhami & Ayton, 2001). Exploring such differences is an important part of understanding judicial decision-making, and nomothetic research cannot examine such inter- or intra-judge variation in the same way as idiographic research can (multi-level models can be used to analyze judge-level differences but cannot determine individual decision strategies; see Dhami & Belton, 2016).

Less commonly, but more appropriately, some researchers have followed the idiographic tradition and examined the decision-making of individual judges making decisions over multiple cases (e.g., Konečni & Ebbesen, 1984; Sensibaugh & Allegeier, 1996). However, when analyzing the data, they use correlational and predictive (regression) statistics, and often appear to treat these models as an isomorphic representation of the judicial decision-making process (i.e., as if they actually represent what goes on within a judge's mind). For instance, when reporting the results of a study of judges' sentencing decisions, Konečni and Ebbesen (1984, p.13) concluded that "Multiple regression and causal analyses revealed that in almost 90% of the cases the judge was responding directly to only one factor". The findings of studies using such statistical models lack relevance because these models are psychologically implausible – judges cannot compute these statistics in their heads. Consequently, computationally complex statistical models are "an unrealistic description of how people make decisions" (Marewski, Gaissmaier, & Gigerenzer, 2010, p. 105). Although as Hoffman (1960, p. 125) points out, all models are only a

"paramorphic representation" of the judgment process, he also warns that the use of statistical models means "one cannot conclude that the mental process has been 'discovered'" (p. 124). We discuss issues of psychological plausibility in more detail later.

In the next two sections, we describe how two approaches developed by researchers investigating the psychology of human judgment and decision-making generally can be used to obtain and analyze judicial decision-making data that are potentially more valid and generalizable. These two approaches are representative design and use of non-statistical, more psychologically plausible models of decision-making.

Collect Judicial Decision Data Using Representative Design

The psychological research carried out by Egon Brunswik offers a methodology for collecting data that better reflects judges' real world decision-making. Brunswik asserted that psychological processes are adapted to the environments in which they function (Brunswik, 1952, 1955, 1956). In his theory of probabilistic functionalism, Brunswik (1952) described the nature of this adaptation which is illustrated in Figure 1. Here, in order to achieve a distal criterion, individuals must learn to infer it from proximal cues.[1] Decision environments are typically probabilistic as cues are only uncertain indicators. These environments may present opportunities for inter-substituting interrelated cues. Thus, when studying psychological processes, researchers ought to use stimuli that are representative of the environments to participants have learned to respond. Brunswik called this 'representative design' (see Dhami, Hertwig & Hoffrage, 2004).

Insert Figure 1 about here

The stimuli presented to participants should be representative of the decision environment in terms of the nature and number of cues (e.g., defendant age, seriousness of offence), their values (e.g., defendant age may be divided into young = 20-29, middle-aged = 30-

59, and old = 60+), distributions (e.g., twice more young than old defendants), co-variation (e.g., defendant age negatively related to defendant health such that old defendants are in poorer health), and ecological validities (e.g., when making sentencing decisions in sex offence cases, a defendant's age and health may be related to sentence severity). Brunswik (1955) suggested that if studying the full population of stimuli is impossible or impractical, researchers should randomly sample stimuli from a defined population or reference class of stimuli from the participant's decision environment, and to which the researcher wants to generalize the findings. Thus, representative design also embraces the idiographic tradition, where each participant is presented with multiple stimuli and his/her behavior is measured over a series of responses.

Figure 1 depicts the possible dependent variables that can be measured. Of particular relevance to researchers studying judicial decision-making are the weights (referred to as the 'cue utilization validities') that an individual attaches to cues (i.e., how important he/she considers the various facts in a case to be). As we will discuss below, representative design enables researchers to examine how an individual perceives and responds to his/her environment, and his/her level of achievement (Brunswik, 1952). Findings can therefore be considered reliable and valid because they capture the individual's natural response patterns. Representative design also enables generalizability of the findings beyond the laboratory or research context (Brunswik, 1956). The findings can be used to predict an individual's future behavior in the environment studied. We will discuss this further in relation to other research designs.

In order to make representative design more practical, Hammond (1966) offered an alternative to conducting research in the field by distinguishing between substantive and formal situational sampling. Substantive situational sampling refers to Brunswik's original proposal to sample real cases from a population. Formal situational sampling, by contrast, refers to the idea

that hypothetical stimuli can be constructed to be representative of real cases in a population in terms of cues, their values, distributions, inter-correlations and ecological validities (see Hammond, Hamm, Grassia, & Pearson, 1987).

Analytically, the lack of experimental control means that variables will naturally co-vary. Brunswik's (1956) solution was to deal with co-variation at the analysis stage of research through, for example, partial correlations. The extent to which the problem of co-variation applies to specific judicial decision-making environments remains to be known. Dhami (2003) found that the average correlations among cues in bail hearings in two English courts were only .2 and .1, respectively.

In order to draw a random sample of real stimuli as Brunswik proposed or to conduct formal situational sampling as Hammond suggested, the researcher needs to define a population or reference class of stimuli. This can be difficult, and while we cannot offer a complete solution to this problem, we suggest considering the environment as that which is subjectively interpreted by the individual. Thus, in a study on judicial decision-making, the reference class is determined following an examination of how a judge perceives and interprets the information relating to a specific decision task, for example, by performing a 'task analysis' (Cooksey, 1996).[2] This could also include establishing the historical experience a judge has with the task. As Brunswik (1944) noted, an individual may approach the same task from different internal states (e.g., attitudes, motivations or emotions), and these could also be considered when defining the reference class.

In acknowledging that representatively designed studies may not fully replace systematically designed studies, Brunswik (1944, p. 37) argued that at the very least representatively designed studies should be used as a type of "check-up" to assess the "soundness" of a systematically designed experiment. For example, in the judicial decision-

making context, Dhami (2003) conducted a four-month observational study of bail hearings in two criminal courts inside and outside London, UK. The details of 342 hearings including the presence or absence of 25 cues (e.g., gender, community ties, and previous convictions), and the decisions made by benches of judges were recorded. Each court's bail decisions were then analyzed separately.[3] The study identified that bail decisions in each court were best predicted by models that included only three cues, and based a decision on only one of these in each case. This more representatively designed study verified the findings from an earlier study of individual judges' bail decisions made on hypothetical cases designed via a fractional-factorial combination of nine cues, that suggested such decisions were made using a simple, non-compensatory strategy (Dhami & Ayton, 2001).

It is important that researchers studying judicial decision-making add representative design to their methodological toolbox because evidence suggests that design matters (see Dhami et al., 2004). For instance, research findings differ depending on how the stimuli (cases) presented to participants were designed. Hammond and Stewart (1974) trained participants to learn to achieve in a simple, two-cue environment, and found that the group of participants later presented with stimuli representative of this environment were more likely to use the cues in a linear way compared to the group later presented with stimuli comprising a factorial combination of the two cues. Others have found that participants presented with hypothetical stimuli designed using factorial combinations of cues used more cues or attached different weights to cues than participants presented with more realistic stimuli (Moore & Holbrook, 1990; Phelps & Shanteau, 1978). Phelps and Shanteau (1978) proposed that the natural inter-correlation among variables may be at least partially responsible for the difference in judgment policies observed under different conditions.

The results of research demonstrating cognitive biases also appear to be affected by representative versus systematic stimulus sampling. For example, researchers have shown that the overconfidence effect (i.e., how well people are calibrated with the accuracy of their knowledge), occurs because past researchers did not sample general-knowledge questions randomly but instead systematically selected hard questions (e.g., Gigerenzer, Hoffrage, & Kleinbölting, 1991). Similarly, Winman (1997) found that hindsight bias or the "I knew it all along" effect was prominent in general-knowledge questions involving paired-comparison tasks when the individual items in each pair were systematically selected rather than when they were representatively selected (i.e., drawn randomly from a specified reference class).

In sum, critics of judicial decision-making research may argue that findings lack internal and external validity because researchers did not study judges making decisions on real cases. The concern is with the lack of realism or representativeness of the case stimuli presented to judges participating in research. In the future, researchers can respond to this by employing representative design when collecting data on judicial decisions.

Analyze Individual Judges' Decision Data Using Psychologically Plausible Models

Selective perception, attention, sequential processing, limited computational ability, and limited memory all have implications for human decision-making. Under these circumstances, people may use decision-making strategies that reduce cognitive effort (e.g., Simon, 1956). In addition, the structure and demands of the decision task such as amount of information available and its redundancy, information presentation format and order, number of response options, and time pressure also influence how decisions are made. According to Hammond's (1996, 2000) cognitive continuum theory, some of these task properties may lead people to abandon analytic thought and move to quasi-rational or intuitive cognition (see Dhami & Thomson, 2012; see also

Dhami, Belton, & Goodman-Delahunty, 2015). Empirical evidence supports these claims (e.g., Dunwoody, Haarbauer, Mahan, Marino, & Tang, 2000; Hamm, 1988; Hammond et al., 1987). The depiction of human decision-making by the statistical models used to analyze data from decision-making studies, however, is incompatible with these considerations.

Since the mid-90s, researchers in the field of judgment and decision-making have argued that non-statistical, more psychologically plausible models ought to be used to capture the decision-making processes of individuals (e.g., Dhami & Harries, 2001; Gigerenzer & Goldstein, 1996; Gigerenzer, Todd & the ABC Group, 1999), including court judges (Dhami, 2003, Dhami & Ayton, 2001; von Helversen & Rieskamp, 2009). In particular, these researchers have proposed the use of simple process models called 'fast and frugal' heuristics as an alternative to past researchers' over reliance on statistical (regression) models.

As stated earlier, regression models are psychologically implausible. Beyond the complex statistical calculations involved, they depict an individual searching through all of the available information, and weighting and integrating it in a compensatory way when making a decision.[4] Furthermore, these models provide only a static view of decision-making, as they suggest that the same cues are used in the same way in every case (Dhami & Harries, 2001). By contrast, simple heuristics are step-by-step process models that embody principles for information search, stop, and decision-making.[5] For instance, cues may be searched in a specific order or randomly, and search may stop when the first cue supporting a particular decision is found (or if out of time). The decision-making process is considered to be fast and frugal as the heuristics search and use little information in a short period of time. Indeed, several simple heuristics are non-compensatory as they base decisions on one cue alone, regardless of how many cues are searched.

These models also imply that individuals make decisions in a flexible manner (i.e., different cues can be used to make decisions on different cases).

Simple heuristics have been developed for various types of decision tasks including two-alternative choice tasks (Gigerenzer & Goldstein, 1996) and binary classification tasks (e.g., Dhami & Ayton, 2001). The 'matching heuristic' is a simple heuristic that has been successfully applied to judicial decision-making (Dhami, 2003; Dhami & Ayton, 2001). This is used for binary classification tasks, and Figure 2 shows an example of the heuristic where the maximum number of cues searched is two. In the context of sentencing an assault case, imagine if the focal decision was custody while the default was community sentence, and the two cues might be seriousness of offence and previous convictions. When presented with a case, the heuristic assumes the judge searches these two cues in order for reasons (or critical cue values) to give a custodial sentence. If a critical cue value is found for a cue that is searched (e.g., the offence is 'very' serious) then search is stopped and the heuristic predicts the judge will pass a custodial sentence. If not, then search continues until the last cue is searched, and if by this time none of the cues searched have a critical cue value, then the heuristic predicts the judge will make the default decision to pass a community sentence.[6]

Insert Figure 2 about here

The matching heuristic is derived on the basis of the relations between cues and decisions over a set of cases and so studies should present each judge with multiple cases, recording the cues available in these cases and the decisions made (see Table 5 in the Appendix to Snook, Dhami and Kavanagh, 2011, for full details of how the heuristic is computed). The critical cue value (reason to make a focal decision) for each cue is defined as the value of the cue that was most frequently assigned a focal decision (e.g., the critical cue value for the previous convictions

cue is 'has priors' if more cases with priors were given a custodial sentence than those without priors). The order of information search is by the rank order of cues as determined by their utilization validities: A cue's utilization validity is the proportion of cases with the critical value that were assigned a focal decision (e.g., the proportion of cases with priors that were given a custodial sentence). Finally, the maximum number of cues searched by the heuristic (*K*) is determined by testing the fit of the heuristic with $K = N$ cues, $K = N$-1 cues and so forth, until the *K* with best fit (i.e. highest percentage of decisions predicted correctly) is identified.

The matching heuristic characterizes information search as lexicographic (ordered according to rank of cues determined by their utilization validities), and cue use as non-compensatory since the decision is based on the value of one cue alone. It defines stopping rules in terms of particular (critical) values of cues, and distinguishes between a focal and a default decision. The psychological plausibility of the matching heuristic does not just lie in the fact that it is non-statistical and non-compensatory, but also in how it assumes individuals learn to use the strategy.[7] It uses frequencies, which Gigerenzer and Hoffrage (1995) argue are a natural form of processing since people (and other animals) seem to learn about contingencies through "natural sampling", "sequential encoding and updating of event frequencies" in their environment (p. 686). The heuristic also involves matching individual cases to a prototype that may be in the mind of a judge, and this is consistent with research on categorization (Estes, 1994). Finally, the matching heuristic exploits cues that may be acquired through direct experience in the domain. The means by which this is accomplished are likely to depend on the sort of mechanisms that Nisbett and Ross (1980) and others have identified as underlying the process of learning about causation and co-variation. The critical value embodies a type of positive-test bias in which only the information that indicates a focal decision is searched and used. There is general evidence for

such strategies in other domains (e.g., Klayman & Ha, 1987). These mechanisms are encapsulated in calculation of cue-utilization validities.

Studies have shown that the matching heuristic is as good as regression and non-statistical linear (compensatory) models in fitting decision data in expert or professional decision-making domains such as medicine, law and crime, and that it outperforms these compensatory models at cross-validation where it makes predictions on a new sample of data (e.g., Dhami & Harries, 2001; Dhami & Harries, 2010; Kee et al., 2003; Snook et al., 2011). In the judicial decision-making context, Dhami and Ayton (2001) found that when predicting individual judges' bail decisions on hypothetical cases comprising nine cues, the matching heuristic contained only one cue for 75% of the judges. When testing the power of the matching heuristic in predicting bail decisions against two additive (compensatory) strategies that used all nine cues (i.e., Franklin's rule which differentially weighted the cues and Dawes' rule which weighted cues equally), it was found that the matching heuristic performed better than the other two models (i.e., 66%, 59% and 63% of decisions predicted correctly by the matching heuristic, Franklin's rule and Dawes' rule, respectively). The heuristic also proved to be the best fit for a greater proportion of the judges.

In a study predicting bail decisions made on real cases by benches (groups of judges) in two courts, Dhami (2003) found that the matching heuristic (containing only three out of a possible 25 cues available) correctly predicted, on average, 92% and 85% of decisions in court A and B, respectively. This was compared to the 86% and 73% of decisions predicted correctly in court A and B, respectively, by Franklin's rule which contained all 25 cues.[8]

In sum, it can be argued that the findings of most research into judicial decision-making lack relevance to any specific judge because researchers have analyzed data aggregated over

groups of judges and/or used psychologically implausible statistical models. In future, researchers can respond to this by applying more psychologically plausible models to the decision-making of individual judges, benches or courts.

Conclusions and Implications

Judicial decisions such as sentencing in a criminal case or awarding damages in a civil case can have huge ramifications for the individuals involved (e.g., loss of liberty and financial loss, respectively). In addition, these decisions can affect the social fabric of wider society which is partly bound together by perceptions of justice and fairness as well as a desire to punish wrongdoers and repair the harm done. Those who administer the justice system are also affected by judicial decisions (e.g., custodial sentences can have resource implications for the prison system). It is no surprise therefore, that judicial decisions come under great scrutiny.

However, to-date, the judicial community and policy-makers in the areas of criminal and civil justice have generally managed to resist making changes to existing policies and practices based on the empirical findings obtained by researchers studying judicial decision-making. Judges have been able to argue, with some force, that for reasons such as those described earlier, the findings are invalid and/or not generalizable to the real world, that they tell us nothing useful or meaningful about how individual judges really do their job. The proposals we present in this article should help researchers in the future to rebut those arguments.

We argue that ideally researchers wishing to understand how judges make their decisions should use representative design to collect data on judicial decisions and then analyze the data using psychologically plausible models. If decisions are made by individual judges, then study them individually. If decisions are made by benches, then study them as benches. Collect decision data across multiple cases. If one cannot ideally study decisions made on real cases in

real time, then sample cases representatively from those the judge/bench has already decided on. If this is not possible, then construct simulated/hypothetical cases that are representative of those the judge/bench normally would come across. Analyze the data by individual judge/bench, using psychologically plausible models for the specific decision-making task at hand.

This calls for a shift away from commonly accepted research practices in disciplines such as psychology, criminology and law. At the very least, researchers should use a triangulated approach to studying judicial decision-making (e.g., using systematic and representative research designs, and compensatory and non-compensatory analytic models).[9] Using the approach proposed here should ultimately allow researchers to learn more about the legal environments and judges that are the topic of their studies. Indeed, representative design requires that researchers understand the parameters of the decision environment in which a judge operates (e.g., cues, their ranges, distributions and combinations). Studies that additionally examine judicial decisions made under different relevant internal and external states (e.g., attitudes and time pressure, respectively) can provide further insights. In addition, the use of more psychologically plausible models requires that researchers understand human cognition and how it functions under different conditions. There is already a large body of knowledge that can be drawn upon, as well as existing models that can be applied (see Dhami, Schlottmann, & Waldmann, 2011; Gigerenzer, Hertwig, & Pachur, 2011).

The proposed changes to research practices in studies of judicial decision-making would allow researchers to generate both internally and externally valid findings that can illuminate whether judges' decision strategies follow that depicted by the scales of justice metaphor. For example, are judges using the legally relevant information? Are they giving the legally relevant information more weight than extra-legal information? Are they integrating the legally relevant

information? An idiographic approach would additionally allow analysis of inter-judge variations in decision strategy. It would be much more difficult for judges and legal policy-makers to dismiss data of this kind as being irrelevant and so it could have greater impact than past research has had to-date.

A meta-analytic view of research methodology highlights that how we study a phenomenon of interest has an impact on our research findings. It remains to be seen to what extent judges fall prey to a myriad of cognitive illusions that can bias their decisions, when they are actually tested under representative conditions. In an historical review of methodology, Gigerenzer and Murray (1987) observed evidence for a "tools-to-theories" hypothesis, where a "scientist's tools…lend themselves to transformation into metaphors of mind" (p. 3). When more psychologically plausible models are tested, the findings dispel the myth of judicial decision-making as a cognitive balancing act. The scales of justice represent a normative ideal that does not fit with psychological reality.

Only when we can paint a valid and generalizable picture of judicial decision-making can we then attempt to understand why judicial behavior departs from the normative ideal. Once we can explain these departures, we can then try to develop interventions to improve judicial decision-making. Judicial decision-making research has not always gone beyond descriptive aims to explanation and intervention. Thus, there is much to do.

References

Baumer, E. P., Messner, S. F., & Felson, R. B. (2006). The role of victim characteristics in the disposition of murder cases. *Justice Quarterly, 17*(2), 281-307. DOI: 10.1080/07418820000096331

Brunswik, E. (1944). Distal focussing of perception: Size constancy in a representative sample of situations. *Psychological Monographs, 56*(1), 1-49. DOI: 10.1037/h0093505

Brunswik, E. (1952). The conceptual framework of psychology. In *International encyclopedia of unified science* (Vol. 1, No. 10, pp. 656-760). Chicago, IL: University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193-217. DOI: 10.1037/h0047470

Brunswik, E. (1956). *Perception and the representative design of experiments*. Berkeley, CA: University of California Press.

Castro-Rodrigues, A. de, & Sacao, A. (2012). Letting the field show us the way – a mixed methodology to understand judicial decision making. *International Journal of Applied Psychology, 2*, 92-97. DOI: 10.5923/j.ijap.20120205.03

Chew, P. K., & Kelley, R. E. (2008). Myth of the color-blind judge: An empirical analysis of racial harassment cases. *Washington University Law Review, 86*, 1117-1166.

Cohen, T. H. (2006). *Appeals from general civil trials in 46 large counties, 2001-2005*. US Department of Justice. Retrieved from http://www.bjs.gov/content/pub/pdf/agctlc05.pdf

Collins, P. M. (2008). The consistency of judicial choice. *The Journal of Politics, 70*, 861-873.

Cooksey, R.W. (1996). The methodology of social judgment theory. *Thinking and Reasoning, 2*(2/3), 141-173. DOI: 10.1080/135467896394483

Coontz, P. (2000). Gender and judicial decisions: Do female judges decide cases differently than male judges? *Gender Issues, 18,* 59-73. DOI: 10.1017/S002238160808081X

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors for judicial decision-making. *PNAS, 108,* 6889-6892. DOI: 10.1073/pnas.1018033108

Dhami, M. K. (2003). Psychological models of professional decision-making. *Psychological Science, 14,* 175-180. DOI: 10.1111/1467-9280.01438

Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making, 14,* 141-168. DOI:10.1002/bdm.371

Dhami, M. K., & Belton, I. (2016). Statistical analyses of court decisions: The example of multilevel models of sentencing. *Law and Method, October 2016.* DOI: 10.5553/REM/.000019

Dhami, M. K., & Belton, I. (2015). Using court records for sentencing research: Pitfalls and possibilities. In J. Roberts (Ed.), *Exploring sentencing in England and Wales* (pp. 18-34). Basingstoke, Hampshire: Palgrave Macmillan.

Dhami, M. K., Belton, I., & Goodman-Delahunty, J. (2015). Quasi-rational models of sentencing. *Journal of Applied Research on Memory and Cognition, 4,* 239-247. DOI: 10.1016/j.jarmac.2014.07.009

Dhami, M. K. & Harries, C. (2001). Fast and frugal versus regression models of human judgment. *Thinking & Reasoning, 7,* 5-27. DOI: 10.1080/13546780042000019

Dhami, M. K., & Harries, C. (2010). Information search in heuristic decision making. *Applied Cognitive Psychology, 24, 571-586*. DOI: 10.1002/acp.1575

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959-988. DOI: 10.1037/0033-2909.130.6.959

Dhami, M. K., Schlottmann, A., & Waldmann, M. (Eds.), (2011). *Judgment and decision making as a skill: Learning, development, and evolution*. Cambridge: Cambridge University Press.

Dhami, M. K., & Thomson, M. (2012). On the relevance of cognitive continuum theory for understanding management judgment and decision making. *European Management Journal, 30*, 316-326. DOI: 10.1016/j.emj.2012.02.002

Doerner, J. K., & Demuth, S. (2010). The independent and joint effects of race, gender, and age on sentencing outcomes in U.S. federal courts. *Justice Quarterly, 27*(1), 1-27. DOI: 10.1080/07418820902926197

Dunwoody, P., Haarbauer, E., Mahan, R., Marino, C., & Tang, C. (2000). Cognitive adaptation and its consequences: A test of Cognitive Continuum Theory. *Journal of Behavioral Decision Making, 13*, 35–54. DOI: 10.1002/(SICI)1099-0771(200001/03)13:1<35::AID-BDM339>3.0.CO;2-U

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin, 32*, 188-200. DOI: 10.1177/0146167205282152

Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650-669. DOI: 10.1037/0033-295X.103.4.650

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.), (2011). *Heuristics: The foundations of adaptive Behavior*. New York: Oxford University Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review, 102*(4), 684-704. DOI: 10.1037/0033-295X.102.4.684

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506-528. DOI: 10.1037/0033-295X.98.4.506

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (Eds.), (1999). *Simple heuristics that make us smart* (pp. 3-34). New York: Oxford University Press.

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2001). Inside the judicial mind. *Cornell Law Review, 86,* 777-830.

Hamm, R. M. (1988). Moment-by-moment variation in experts' analytic and intuitive cognitive activity. *IEEE Transactions on Systems, Man, and Cybernetics, 18,* 757–776. DOI: 10.1109/21.21602

Hammond, K. R. (1966). Probabilistic functionalism: Egon Brunswik's integration of the history, theory, and method of psychology. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 15-80). New York: Holt, Rinehart and Winston.

Hammond, K. R. (1996). *Human judgment and social policy: irreducible uncertainty, inevitable error, unavoidable injustice*, Oxford, England: Oxford University Press,

Hammond, K. R. (2000). *Judgments under stress.* New York: Oxford University Press..

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytic cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics, 17,* 753-770. DOI: 10.1109/TSMC.1987.6499282

Hammond, K. R., & Stewart, T. R. (1974). *The interaction between design and discovery in the study of human judgment.* Report No. 152, University of Colorado, Institute of Behavioral Science.

Harris, J. C., & Jesilow, P. (2000). It's not the old ball game: Three strikes and the courtroom workgroup. *Justice Quarterly, 17*(1), 185-203. DOI: 10.1080/07418820000094521

Hastie, R., & Viscusi, K. (1998). What juries can't do well: The jury's performance as a risk manager. *Arizona Law Review, 40,* 90-921.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 47,* 116-131. DOI: 10.1037/h0047807

Judicial Appointments Commission (2011). *Annual report and accounts 2010/2011*. Retrieved from http://jac.judiciary.gov.uk/static/documents/JAC_Web_cover_2011_Final_New.pdf

Kahneman, D., Slovic, P., & Tversky, A. (Eds.), (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.

Kee, F., Jenkins, J., McIlwaine, S., Patterson, C., Harper, S., & Shields, M. (2003). Fast and frugal models of clinical judgment in novice and expert physicians. *Medical Decision Making, 23,* 293–300. DOI: 10.1177/0272989X03256004

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211-228. DOI: 10.1037/0033-295X.94.2.211

Konečni, V. J., & Ebbesen, E. B. (1984). The mythology of legal decision making. *International Journal of Law and Psychiatry, 7*(1), 5-18. DOI: 10.1016/0160-2527(84)90003-7

Kulick, C. T., Perry, E. L., & Pepper, M. B. (2003). Here comes the judge: The influence of judge personal characteristics on federal sexual harassment case outcomes. *Law and Human Behavior, 27,* 69-86. DOI: 10.1177/2066220315595906

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast and frugal trees. *Psychological Review, 118,* 316-338. DOI: 10.1037/a0022684.

Manning, K. L., Carroll, B. A., & Carp, R. A. (2004). Does age matter? Judicial decision making in age discrimination cases. *Social Science Quarterly, 85,* 1-18. DOI: DOI: 10.1111/j.0038-4941.2004.08501001.

Marewski, J.N., Gaissmaier, W., & Gigerenzer, G. (2010). Good judgments do not require complex cognition. *Cognitive Processing, 11*(2), 103-121. DOI: 10.1007/s10339-009-0337-0

Martin, E., & Pyle, B. (2004). State high courts and divorce: The impact of judicial gender. *University of Toledo Law Review, 36,* 923-947.

Merrall, E. L. C., Dhami, M. K., & Bird, S. M. (2010). Exploring methods to investigate sentencing decisions. *Evaluation Review, 34,* 185–219. DOI: 10.1177/0193841X10369624.

Mitchell, O. (2005). A meta-analysis of race and sentencing research: Explaining the inconsistencies. *Journal of Quantitative Criminology, 21,* 439-466. DOI: 10.1177/0887403409354738

Moore, W. L., & Holbrook, M. B. (1990). Conjoint analysis on objects with environmentally correlated attributes: The questionable importance of representative design. *Journal of Consumer Research, 16,* 490–497. DOI: 10.1086/209234

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences, 37*(1), 1-19. DOI: 10.1017/S0140525X12003214

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review,* 84, 231–59. DOI: 10.1037/0033-295X.84.3.231

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. R. Wrightsman, (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Peresie, J. L. (2004). Female judges matter: Gender and collegial decisionmaking in the federal appellate courts. *Yale Law Journal, 114,* 1759-1790.

Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance,21,* 209–219. DOI: 10.1016/0030-5073(78)90050-8

Rachlinski, J. J., Guthrie, C., & Wistrich, A. J. (2011). Probable cause, probability, and hindsight. *Journal of Empirical Legal Studies, 8,* 72-98. DOI: 10.1111/j.1740-1461.2011.01230.x

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2009). Does unconscious racial bias affect trial judges? *Notre Dame Law Review, 84,* 1196-1246.

Robbennolt, J. (2002). Punitive damage decision-making: The decisions of citizens and trial court judges. *Law and Human Behavior, 26,* 315-341. DOI: 10.1023/A:1015376421813

Sensibaugh, C. C., & Allegeier, E. R. (1996). Factors considered by Ohio juvenile court judges in judicial bypass judgments: A policy-capturing approach. *Politics and Life Sciences, 15,* 35-47.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63,* 129-138. DOI: 10.1037/h0042769

Snook, B., Dhami, M. K., & Kavanagh, J. (2011). Simply criminal: Predicting burglars' occupancy decisions with a simple heuristic. *Law and Human Behavior, 35,* 316-326. DOI: 10.1007/s10979-010-9238-0

Turner, K. B., & Johnson, J. B. (2006). The effect of gender on the judicial pretrial decision of bail amount set. *Federal Probation Journal, 70*, 56-62.

Von Helversen, B., & Rieskamp, J. (2009). Predicting sentencing for low-level crimes: Comparing models of human judgment. *Journal of Experimental Psychology: Applied, 15*(4), 375-395. DOI: 10.1037/a0018024

Wooldredge, J., Griffin, T., & Thistlethwaite, A. (2013). Comparing between-judge disparities in imprisonment decisions across sentencing regimes in Ohio. *The Justice System Journal, 34*, 345-368. DOI: 10.1080/0098261X.2013.10768044

Winman, A. (1997). The importance of item selection in "knew-it-all-along" studies of general knowledge. *Scandinavian Journal of Psychology,38,* 63–72. DOI: 10.1111/1467-9450.00010

Endnotes

[1] The distal criterion is not always obvious in the judicial domain. For instance, some may argue that judges ought to strive for decisions consistent with past ones in similar cases, while others might argue that judges ought to strive for decisions that are perceived to be just or fair.

[2] A task analysis provides information about how a judge perceives his/her decision environment (e.g., number, nature of cues, inter-relations etc). This allows construction of representative stimuli which are then presented to the judge. His/her decision behavior is then measured.

[3] Each court's (rather than bench's) decisions were modeled because benches were not static entities – judges in each court sat on different benches, and a bench made only a few decisions.

[4] Although cue weights may be non-compensatory and non-linear terms may be included, it is generally assumed that judgments are the product of a linear, compensatory integration of multiple cues that are weighted optimally.

[5] These heuristics differ from those characteristic of the "heuristics and biases" research program, which are vague about process and exclude pre-decisional behavior such as information search (see Kahneman, Slovic, & Tversky, 1982).

[6] There are different rules for missing/unavailable cue values depending on the task (Dhami, 2003).

[7] Indeed, not all heuristics are based on simple learning (Luan, Schooler, & Gigerenzer, 2011).

[8] Similarly, von Helversen and Rieskamp (2009) found that a fast and frugal heuristic of quantitative estimation predicted German prosecutors' decisions in real sentencing cases better than a linear regression model or Dawes' rule.

[9] Others have recently also argued that researchers ought to use a triangulated approach that incorporates qualitative data (Castro-Rodrigues & Sacao, 2012).

Figure 1. Adapted lens model. From "The conceptual framework of psychology." In *International Encyclopedia of Unified Science* (p. 678), by E. Brunswik, 1952, Chicago: University of Chicago Press. Copyright 1952 by the University of Chicago Press.

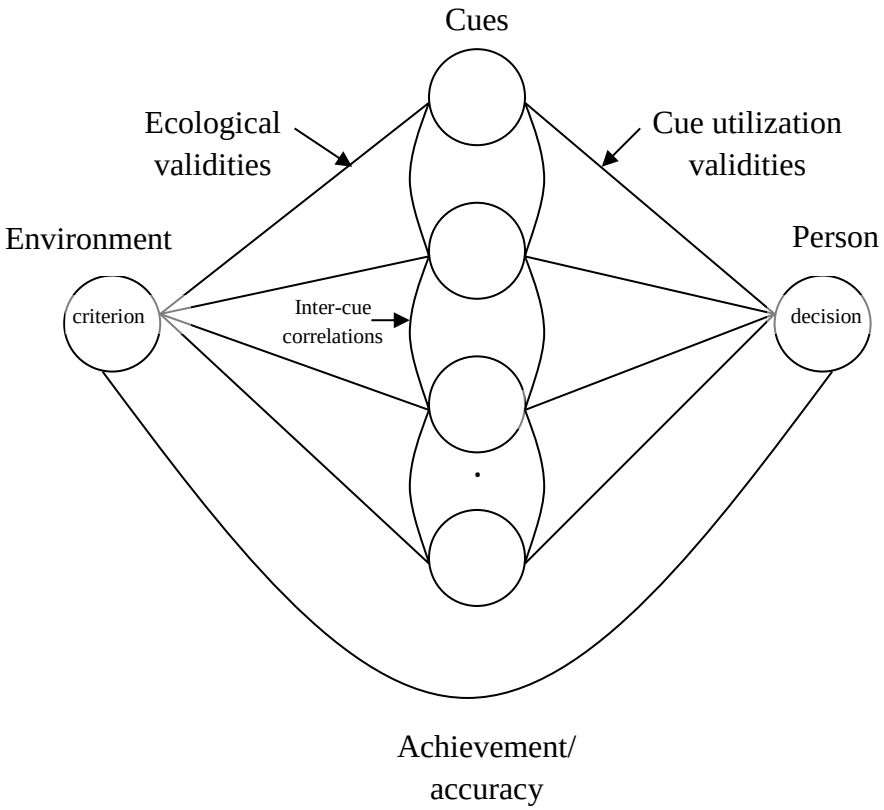

***Figure 2*.** Example of the Matching Heuristic where a maximum of 2 cues are searched.