<https://helda.helsinki.fi>

Predefined Criteria and Interpretative Flexibility in Legal pö Courts Evaluation of Expertise

Taipale, Jaakko

2019-11-01

pö Taipale , J 2019 , ' Predefined Criteria and Interpretative Flexibility i
Evaluation of Expertise ' , Public Understanding of Science , vol. 28 , no. 8 , pp. 883-896 . <https://doi.org/10.1177/0963662519881338>

<http://hdl.handle.net/10138/310753>

<https://doi.org/10.1177/0963662519881338>

unspecified

acceptedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Predefined Criteria and Interpretative Flexibility in Legal Courts' Evaluation of Expertise

Abstract

This study examines two different approaches in empirical analysis of judges' evaluation of expertise in court: firstly, an analyst-based approach that employs predefined normative criteria to measure judges' performance, and secondly, an actor-based approach that emphasizes interpretative flexibility in judges' evaluation practice. I demonstrate how these different approaches to investigating judges' adjudication lead to differing understandings about judges' abilities to evaluate scientific evidence and testimonial. Although the choice of analytical approach might depend on context and purpose in general, I contend that in assessing judges' competence an actor-based approach that adequately describes the way in which judges relate to and handle expertise is required to properly understand and explain how judges evaluate expertise. The choice of approach is especially important if the resulting understanding of judges' competence is subsequently used as a basis for making normative and prescriptive claims with potential consequences for trial outcomes.

Keywords: expertise, interpretative flexibility, scientific literacy, law and science, Daubert, traumatic brain injury, DTI, fMRI, practice

Introduction

The pressing general question concerning efficiency and legitimacy of legal courts is how judges can competently make use of expertise in finding the facts and bringing forth closure in a case (Jasanoff, 2015: 1724). Translated into a question for professional training, how should judges of law be taught about expertise and how to evaluate it in a court of law? I argue that this question is subordinate to a second question examined in this paper: how should we study, understand and explain judges' evaluation of expertise in court? The second question bears directly on the first, because different approaches to studying judges' performance produce differing results, and the difference is highly relevant to the question of how judges should be educated about expertise and its evaluation.

The argument is pursued in the article by examining two different ways to investigate trial judges' competence in evaluating and making decisions based on scientific evidence and testimonial. I start by discussing an analytical approach that employs predefined categories to assess judges'

performance. Conceptually the issue revolves around the so-called Daubert criteria, devised by the US high court in a string of three 1990s decisions. The criteria act as guidelines for US judges' admissibility decisions regarding scientific expertise (see Jasanoff, 2005: 49-50; Edmond, 2002). As discussed by Michael Lynch (2014: 108), the Daubert criteria can be understood as demarcation criteria for telling relevant and reliable expert evidence from its opposite. Thus, the criteria involve a presupposition that an optimal way to evaluate scientific evidence and testimonial exists. In terms of assessing judges' evaluation of expertise, the predefined criteria offer a general yardstick for proper, science-based evaluation of expert evidence and testimonial. In consequence, judges' failure to understand and use the criteria in a way determined as proper is taken to indicate judges' lacking competence in science. This normative line of thought introduces the idea of improving judges' or jurors' scientific literacy to improve courts' performance. It also, at least implicitly, casts fact finders as a source of distortion in the process of communicating relevant scientific evidence between experts and courts. What essentially is the scientific literacy deficit model, as discussed by Edmond and Mercer (1997), seems to be a somewhat shared understanding in psychological studies of jurors' evaluation of expertise in court (e.g. Kovera and McAuliff, 2000: 575-576, 585; Gatowski et al., 2001: 451-455; Wingate and Thornton, 2004: 110-111; Chin, 2014: 227, 234-236; Tadei et al., 2016). Edmond and Mercer (1997: 329, 349-350) have also earlier noted that "most existing approaches measure jury competence against an unproblematic yardstick of the 'correct scientific understanding'". Fast forward two decades, this normative approach lingers on.

As a case in point, I discuss a study that employs the Daubert criteria as a yardstick for measuring Finnish judges' ability to evaluate expertise in court (Tadei et al., 2016). The study is also connected to a Finnish Ministry of Justice advanced training course. In itself, the authors' recruitment of the Daubert criteria is an interesting instance of how generalized and authoritative frameworks of intelligibility transcend cultural boundaries and are adapted in local contexts of professional practice. However, in a more detailed scrutiny the examined study also fails in its central goals, and the related training course might even carry unwanted consequences regarding legal courts' efficiency to deliver justice.

To illustrate how different approaches to assessing judges' performance produce differing understandings of judges' competence in evaluating expertise, I then present my empirical analysis of how judges at the Helsinki district court discuss expertise in a case type that involves highly technical and esoteric expertise in neurology, neuropsychology and neuroradiology. The legal dispute concerns diagnosis of traumatic brain injury (TBI, more specifically diffuse axonal injury (DAI)) and whiplash injuries, and the role of advanced imaging technologies DTI and fMRI in

diagnosing these injuries. I show how judges use criteria similar to Daubert in evaluating expertise, but in a much more flexible and situated way than sticking to general and predefined evaluative criteria would permit: *judges evaluate expertise both by 1) flexibly interpreting evaluative criteria as required by case context and by 2) taking stock of interpretative flexibility in expert evidence and testimonial. It is the situated and pragmatic interplay between the two that characterizes judges' evaluation practice.*

Based on the analysis presented in this paper, I conclude that the predefined criteria approach allows only for a partial view of both judges' competence in and manner of evaluating expertise in court. This results from *descriptive inadequacy* (Israel-Jost and Kinzel, 2014), i.e., not taking account of how examined actors (here judges) themselves define and use the categories that the analysis uses as a set of predefined criteria to measure judges' competence. Thus, relying on predefined criteria will reproduce rather than solve the difficulties related to evaluating expertise. I also contend that the suggested measure of improving courts' performance by improving judges' scientific literacy will not work, because the idea of improving judges' literacy rests on a very general conceptualization of science and a misunderstanding of what expertise is and how it can be acquired.

Therefore, to avoid a reductive account of judges' evaluation practice, we require a descriptive, case-specific and actor-based account of judges' performance, emphasizing *interpretative flexibility* (Baker, 2011; Collins, 1981: 4) as a defining feature of how judges engage expert evidence with evaluative criteria. It is typical of legal disputes that non-expert fact finders (judges) encounter many alternative but equally plausible explanations of evidence. The difficulty, diversity and contested character of knowledge claims requires that judges negotiate this flexibility by sidestepping the all too demanding quest for veracity in the evaluation of expertise, and evaluate experts' and their claims' *credibility* instead. Since the evaluative criteria are difficult to judges much in the same way as expert claims are, the criteria are also subject to interpretative flexibility in how they are used. Understanding and explaining judges' practice of evaluating expert evidence and testimonial as a situated, practical and pragmatic activity should also contribute, if taken into account, to designing more efficient and effective training courses for legal professionals tackling expert evidence in courts. The following conceptual and empirical discussion is relevant also in the context of (e.g.) journalistic practice and administrative decision making, as they share the general problem with judicial decision making, i.e., how can non-experts competently assess the soundness and appropriateness of expertise.

Methodology and data

The study is based on a theoretical distinction between what can be called an *analyst-based* approach that utilizes a *predefined analytical categorization* to investigate social processes and explain agency, and an *actor-based* approach that bases its findings on the manner in which actors that partake in the investigated social processes themselves use such categories to do normative work for them (Lynch, 2014). In this study, the analytical categories derive in both approaches from the aforementioned US high court Daubert decision: *falsifiability*, *error rate*, *peer review and consensus (or scientific/practical acceptance)* (Haack, 2005; Jasanoff, 2005; Tadei, 2016).

My intention is to discuss how different methodological and analytical approaches generate different kinds of understanding about judges' competence in evaluating expertise. I criticize the predefined criteria approach for its inbuilt scientific literacy deficit model and also for generating a reductive account of judges' competence. Edmond and Mercer (1997: 339) argue for constructionist studies which emphasize court interaction as central to de- and reconstruction of knowledge claims, thus explaining the relevance and role that scientific expertise has in courts' decision making more comprehensively than the literacy deficit model does. However, instead of taking the constructionist perspective, I aim to provide a third perspective in examining how single powerful actors, trial judges, manage contested and controversial expert claims in court. I emphasize the relevance of managing alternative but equally plausible explanations of evidence (cf. Edmond and Mercer, 1997: 341), and suggest that the demonstrated *interpretative flexibility* (Baker, 2011; Collins, 1981) is centrally important to describing and explaining how legal courts' evaluate and understand expert evidence and testimonial.

As well as critically discussing the analyst-based, predefined criteria approach both in general and more specifically in the case of assessing Finnish judges' performance (Tadei et al., 2016), the article examines eleven Helsinki district court case verdicts from 2014-2017. The examined case type features a medical dispute concerning brain injury, allegedly received by plaintiffs in a traffic accident. In addition to legal reasoning and technicalities, the eleven verdicts contain long passages that explicate the medical dispute and evidence, various positions taken by the parties and their experts, and judges' views on the disputed issue. The verdicts are approached analytically as performative acts (Law, 2008; cf. also Bartlett et al., 2018: 4-5), in which judges reason for and legitimate their decisions. The 1) contested and inconclusive character of the evidence 2) forces the judges to interpret and choose (and justify the choice) between expert claims that 3) involve scientific (medical and psychological) fields of expertise in which the judges are non-experts. These

three factors produce high interpretative flexibility in the case type. In the verdicts, I analyse actors' (judges') discussion on radiological and neurological expertise, with the purpose of demonstrating the role that interpretative flexibility (Baker, 2011; Collins, 1981) has in how judges engage expertise.

The eleven verdicts, each on average over 40 pages in length, were analysed as part of a larger study that investigates medical expertise in court (Taipale, 2019). The verdicts were coded twice, first with Atlas.ti software, and then by categorizing the once-coded data with the Daubert criteria. Both rounds focused on the topic of MRI technologies.

The actor-analyst distinction is difficult to keep in observing actual social processes. This is particularly so if the categories analysts use as a normative explanative framework are the same categories (e.g. "consensus") that the studied actors use to achieve their preferred ends (Lynch, 2014: 108). A similar point has been made about research processes in general, as the empirical-analytical process unavoidably leads into actors' and analysts' categories being implicated by each other (Collins and Evans, 2014: 786-788). A balanced analysis therefore calls for reflexive iteration between analysts' and actors' concepts and categories. The case-specific difficulty is in knowing what is descriptive enough of the examined actors' practices, while still being meaningful on a more general plane of explanation (cf. Israel-Jost and Kinzel, 2014). Despite these difficulties, the analyst-actor distinction is useful as a descriptive categorization in empirical analyses of cases that lack and would benefit from such reflexivity. The article continues by discussing such a case in detail in the next section.

Describing and explaining with analyst's categories

The study by Tadei et al. (2016) exemplifies well the use of predefined analytical criteria in analysis of agency, and the study also joins the standard prescriptive call to improve judges' scientific literacy (Chin, 2014; Gatowski et al., 2001; Kovera and McAuliff, 2000; Wingate and Thornton, 2004). The study is also the only empirical study on Finnish judges' performance regarding the evaluation of expertise, and at least two of the five authors have been teaching at a Finnish Ministry of Justice (MoJ) advanced training course in psychological expertise and witness psychology. Testing the course participants also yielded data for Tadei et al. (2016). All in all, the examined study potentially has a higher social impact than its scientific impact or quality would suggest: what the authors claim about expertise, judges' relation to science and the best way to engage scientific expertise in court also most likely influences the content of future MoJ training courses.

In the authors' own words, the underlying desire is to gain such an understanding of judicial decision making that would allow determining "how to best train legal professionals to correctly evaluate expert testimony given in court" (Tadei et al., 2016: 13). Resting quite heavily on earlier studies discussing the Daubert criteria, the authors seek to assess what kind of ability judges have in evaluating the reliability of expert claims based on scientific research and practice (Tadei et al., 2016: 2-4, 6, 9, 12). To accomplish this, the authors tested 87 Finnish professionally experienced judges. The judges were instructed to rate the importance of each Daubert criterion, *falsifiability*, *error-rate*, *peer-review*, *scientific acceptance/practical acceptance*, for evaluation of expertise in court. After a brief training about the criteria, the judges were asked to pose questions to imaginary experts in fictitious vignettes. The rated criteria and questions asked were then compared to see if the questions would reflect the rate of importance given to the criteria.

Tadei et al. (2016) found that the 87 judges focused in their evaluation on the auxiliary criterion work experience and practical/scientific acceptance of the evidence in the field of expertise at the expense of the other criteria. "Overall, the judges did not use the criteria a lot when asking questions, even though they had just been presented with the criteria in the first two parts of the study" (Tadei et al., 2016: 10). The main conclusion was that judges might "not have enough knowledge about scientific evidence to properly evaluate the validity of expert testimony" (Tadei et al., 2016: 11), which as a finding sits well with the idea of scientific literacy deficit in related earlier studies I mentioned (e.g. Gatowski et al., 2001; Wingate and Thornton, 2004). The authors concluded that judges do not understand the criteria due to not having received any scientific training. Therefore, the study prescribed scientific training to legal professionals, and stated that instituting guidelines in Finnish code for judicial procedure (CJP) for evaluation of evidence would reduce the probability of "junk science" appearing in court, and also increase transparency in forcing judges to better explain their decisions regarding expert evidence and testimonial.

There are difficulties in Tadei et al. (2016: 11-12) research design and data analysis, which the authors discuss to an admirable extent. Instead of repeating their study, I start from their results and conclusions, and work my way back to their premises – namely the normative research design which employs predefined analytical criteria and the conception of expertise – which both prove problematic, if the goal is to assess, understand and explain judges' practice of evaluating expertise in court, with the intent of then making prescriptions about how to train judges.

The trouble with predefined normative criteria

The difficulties with studies that employ predefined analytical criteria to assess judges' competence and prescribe scientific training to judges arise from, first, the analyst-centred approach to studying judges' performance, and second, from a limited understanding of what expertise is and how it can be acquired. Taken together, the two premises amount to a partial understanding of judges' relation to expertise. I will first discuss the normative approach in the exemplary case Tadei et al. (2016):

The absence of clear guidelines likely diminishes the capacity of judges to evaluate the veracity of expert testimony [...] The decision to investigate the importance [the judges give] to the Daubert criteria, even if these are not used in the Finnish legislation, was made because they are nevertheless *fundamental in addressing the veracity of expert testimony*. A judge who has to evaluate pieces of evidence based on empirical sciences *must know and understand* concepts like falsifiability, error rate, peer-reviewed research, and scientific/practical acceptance. Furthermore, if no more specific guidelines exist, the Daubert criteria present the most clear and precise guidelines for how to relate with science in trials. (Tadei et al., 2016: 4; italics added)

The key concept to understanding the approach is veracity. Evaluating the reliability of expertise and expert claims should focus on their veracity, that is, the correctness and accuracy of presented knowledge compared to the actually existing state of things. To make such judgments of veracity, judges have to understand and competently use scientific principles of evaluating scientific evidence and expert claims, and these principles are captured in the Daubert criteria.

Tadei et al. (2016) consider the Daubert criteria as universal criteria for evaluating expertise, and play down the fact that their study only concerns psychological and psychiatric expertise: "The criteria can, however, be implemented on all expert testimony" (Tadei et al., 2016: 12). It certainly can be implemented, but whether or not the criteria deliver in the evaluation of the veracity of expert claims is debatable (Chin, 2014; Cole and Bertenthal, 2017: 357; Edmond, 2000; Haack, 2005; Jasanoff, 2005). My purpose is not, however, to join the chorus of critical reviews of the mentioned criteria. In keeping with this article's mission, I am more concerned with how the criteria is used as a normative guide, or a yardstick, in evaluating judges' performance. Moreover, as a yardstick, do the criteria actually measure judges' competence in evaluating scientific expertise – or does the criteria really only measure how judges know and understand the criteria itself?

Five of the seven criteria Tadei et al. (2016) used to assess the 87 judges' competence were taken from the Daubert decision and related literature. The Daubert criteria error rate, falsifiability, peer-reviewed research and practical/scientific acceptance were complemented with two additional criterion, work experience and scientific activity, that focus on the expert and his/her professional

credibility rather than the quality of knowledge. Although the authors recognize the difference between the two types of criteria, they do not discuss the difference properly or pursue it in their data, despite predicting that judges would focus on experts' characteristics and overall acceptance of their claims rather than scientific principles. Their results confirmed these expectations, as judges mostly used the social indications of professional experience and general acceptance of expert claims as cues in evaluating the presented evidence (Tadei et al., 2016: 9-11).

Thus, the authors recognize the distinction between credibility and veracity at play in their case. Despite this, their conclusions and prescriptive outcomes are solely based on the evaluation of veracity: "To make the best use of expertise, the quality and veracity of expert testimony should be correctly evaluated by legal experts" (Tadei et al., 2016: 1-2). This normative emphasis is also evident in discussion of limitations: there was "difficulty in coding [the data] in categories, especially into those not included in the original design" (Tadei et al., 2016: 12). The authors speculate that the operationalization of the predefined analytical categories influenced the test group's thin usage of the Daubert criteria. For example, the analysis showed that judges did not really use the criterion falsifiability at all, and the authors suggest it had something to do with how they themselves (the authors/analysts) defined both the category and how judges should meet the category in their questioning of expert witnesses. Regardless of these difficulties, the authors keep to their normative framework: "Daubert (at 593) explicitly refers to the concept of falsifiability, so it should be used by judges who want to properly deal with science in court" (Tadei et al., 2016: 12).

I contend that there is another way of understanding the results of Tadei et al. (2016), which is not related to flaws in study design, and which complicates their conclusions (judges are poor at evaluating scientific evidence and testimonial) and prescriptions (judges should be trained in science, guidelines should be instituted in Finnish CJP). As I have shown, the authors' normative approach entails using predefined criteria (Daubert's scientific ideals) as a yardstick for measuring judges' capacity to evaluate scientific expertise (cf. Edmond and Mercer, 1997). I argue, however, that the predefined criteria are external or even blind to judges' actual practice of evaluating expertise. Therefore using *the criteria as a yardstick results in a very partial understanding of judges' capacities to evaluate expertise*. Such an assessment of judges' competence should instead be guided by a more actor-centred understanding of judges' evaluation practices.

Limited understanding of expertise

For the mentioned psychological studies (Kovera and McAuliff, 2000; Gatowski et al., 2001; Wingate and Thornton, 2004; Chin, 2014; Tadei et al., 2016), the scientific literacy deficit model holds that analysing veracity requires judges to analyse expertise in the scientific domain with scientific principles, comparative to what scientists do when they critically evaluate other research groups' claims and findings. In Tadei et al. (2016) the authors notice the discrepancy between what they expect from the judges, and how judges can reasonably be expected to perform in the light of their current training, but fail to appreciate the depth of the issue. This is evident from the fact that the authors prescribe scientific training to judges. Although an earlier study (Kovera and McAuliff, 2000) indicates that giving judges scientific training might in some cases improve judges' ability to use scientific principles in evaluating expertise, I contend that the very idea of making professional actors (judges) literate in science (cf. Bauer et al., 2007: 80-82) rests on a misunderstanding of what expertise is.

To analyse the veracity of expertise and expert claims requires, in effect, expertise in the analysed field. However, expertise is acquired through a long-term social immersion in an expert community of practitioners. It entails soaking up tacit knowledge about how to do, value and weigh all the mundane practical tasks and problem solving that build up into a trained expert judgment. In addition, expert judgments on scientific claims often require social understanding about personal qualities of other scientists and their standings in the assessed field of expertise (Collins, 1985; Collins and Evans, 2007; Daston and Galison, 2007: 309-321; Fleck, 1935). Consequently, one cannot acquire the required expertise to evaluate the veracity of expert claims from reading textbooks or attending training courses (Haack, 2005: 70).

Expertise is, in other words, socially and practically embedded, and it is always expertise in something in particular (Lynch, 2014: 110). Thus, its correct evaluation does not come about through mastering general criteria of evaluating scientific claims, such as those captured in the Daubert criteria: there is no general expertise, because gaining expertise presupposes experience and tacit knowledge transfer through interacting in a specific expert community. Judges cannot master another specific field of expertise in depth, even if only for the lack of resources such as time. Therefore, training judges how to evaluate scientific expertise by teaching them general evaluation criteria (or instituting criteria as guidelines in the CJP) might result in judges learning the criteria, but they would not become more competent in evaluating the veracity of expert claims.

In addition, no textbooks deal with knowledge in fields of technical expertise that examine the limits of what is known to us. This research frontier (Cole, 1992: 15) produces the most controversial type of scientific evidence that often ends up in courtroom as novel evidence in both new and already established case types. This is the very situation with advanced MRI technologies DTI and fMRI in Helsinki district court TBI cases that I discuss in the next section. The claims the experts in neurology and neuroradiology make in court come supported by tens of years of clinical and medical scientific experience. Taking stock of all this, judges would need a truly magical handbook or training course to help them, directly and by themselves, determine the veracity of DTI and fMRI findings. Nor does the fact that neurological experts themselves are locked in a controversy about the relevance and reliability of their findings support the conclusion that scientific training would help judges. Closure cannot depend on the best facts only – because in the controversy surrounding DTI and fMRI (or in frontier science in general) no one necessarily possesses the facts in the (textbook) sense of having final knowledge.

As Finnish high court judge Kirsti Uusitalo (2013: 13) points out, training judges to use predefined evaluation criteria might also create adverse results. Judges have no scientific training and expertise in the sense described above, and thus have no possible way to fully understand how the evaluation criteria might be used to reveal the veracity of expert claims. This might lead to haphazard or superfluous application of the criteria. Thus, the criteria could be reduced to a legitimization mechanism, instead of offering a set of tools to find the facts in a given case (cf. Gatowski et al., 2001: 453). This amounts to cloaking the problem of expertise in courts rather than solving it.

Because judges lack scientific training and expertise in the deep sense as defined above, judges take a pragmatic approach to controlling the epistemic uncertainty caused by inconclusive and contested expert evidence (Taipale, 2019). This uncertainty can also be conceptualized as high prevalence of interpretative flexibility in the presented evidence. Pragmatism and flexibility also apply to the use of the Daubert criteria. Such use of the criteria can result in a practice that can analogously be understood as a mixed language or grammar rules (cf. Collins, 2011), used and followed by actors to legitimate their statements with regard to their community of practice and audiences external to it (Turner, 2014: 23-26; Kritzer, 2007: 334). However, occasions of such creole language generation (cf. Collins et al., 2007) between two epistemic cultures (i.e. law and science) are also vulnerable to embedding and normalizing of biased or even severely compromised understanding of scientific expertise, expert claims and their evaluation, with potentially adverse consequences considering the contextually specific ends of that evaluation.

Training judges to operate with predefined evaluative criteria and a limited understanding of expertise is one step towards embedding such a compromised understanding into the legal system. Instituting general evaluation criteria into a Finnish code for judicial procedure (CJP) is another. The fact that Finnish legal literature mentions positively the Daubert criteria indicates that the recruitment of the criteria is not a completely foreign idea to the Finnish judiciary (Väisänen and Korkman, 2014:725; Loiva, 2012: 646, Rask, 2011: 25-27). I contend that instead of normative criteria, what we require to improve judges' performance in evaluating expertise is a descriptively adequate, actor-based understanding of judges' evaluation practice (cf. Israel-Jost and Kinzel, 2014; Lynch, 2014).

Interpretative flexibility in Helsinki district court TBI cases

In the following, I examine how judges in court discuss the same or similar criteria to the Daubert criteria, although I mostly focus on the criterion of falsifiability to drive home my point. Thus, the Daubert principles are investigated as actors' categories, with which actors order the world and make sense of their situation. The analysis concludes that *evaluating expertise entails considerable interpretative flexibility in the use of evaluative criteria. I also highlight how judges display a discursive ability to engage the interpretative flexibility in presented expert evidence and testimonial.*

The case type and verdicts I examine concern insurance compensation disputes between individual plaintiffs and insurance companies. The difficulty lies in the injury type that is the basis of compensation claims: the plaintiffs have (allegedly) suffered traumatic brain injuries (TBI) and/or neck injuries in traffic accidents, leaving them disabled. Especially borderline diffuse axonal injuries (DAI) are difficult to diagnose, and diagnoses are mostly based on observable symptoms determined by neurological and neuropsychological testing. The time lapse between the accident and injury diagnosis, and subsequent litigation process, is in most cases considerable. The temporal distance adds to the difficulty of proving both the existence of injury and causation between the accident event and plaintiffs' disabilities.

These difficulties are also the reason why advanced MRI technologies (such as DTI in the case of DAI, and fMRI in the case of cervical vertebrae injury) are so prominently discussed and contested in court. Imaging technologies hold great promise (Jones et al., 2013) in providing tangible physiological evidence that answers the question of whether or not the plaintiff presently suffers from brain or neck injuries, and subsequently, together with likely cause from a given accident, whether the defendant insurance company is liable to compensation. At the same time, both DTI

and fMRI are seen as experimental technologies wrought with problems. The potential medico-legal abuse of such technologies has also been noted (Wortzel et al., 2014). Moreover, how judges handle DTI and fMRI evidence is a suitable object of analysis for this paper, since the examination methods represent highly technical expertise and advanced technology. Competent use of these imaging examination methods requires a great deal of expertise that results from extensive education, training and experience of working in relevant professional fields, as do judgments based on the clinical findings the discussed technologies produce. Despite all the featured experts regarding the presented findings being very experienced and highly credentialed, their opinions vary a great deal. Thus, the field of TBI diagnosis and care is medically controversial, which adds difficulty to judges' fact finding task.

Echoing Gatowski et al. (2001: 453) and Tadei et al. (2016), of all the four Daubert criterion it was hardest to detect discussion relating to the criterion *falsifiability* in the analysed verdicts. However, this is a matter of definitions and contextual understanding; or a matter of identifying the interpretative flexibility of the criterion. Falsifiability in its Popperian-inferred (see Haack, 2005) coinage means that a theoretical proposition can be demarcated as belonging to science if it can, in principle, be proven to be false by means of empirical scientific inquiry. In the TBI cases neither of the parties make outlandish medical claims or present outright charlatan experts to prop them up: both sides present credible medical scientific arguments. Therefore, making a crude demarcation between science and non-science translates into making much finer distinctions in quality of medical practice, i.e., between proper and improper medical and scientific practice.

The theoretical proposition to be proven false in the TBI cases is the following: plaintiff's impairment/symptoms are explained by TBI/neck injuries that result from a given traffic accident. For the purposes of this paper, I concentrate on the central discussion that concerns the possibility of making a reliable diagnosis (or achieving one altogether), and the uses of advanced MRI technologies in making that diagnosis. In addition, falsifiability is defined in Tadei et al. (2016: 6) and also in the Daubert criteria (Haack, 2005: 66-67) as "testability" of theoretical propositions. Such "testability" of MRI results is also discussed in the investigated TBI case verdicts. In this discussion, the proper application of DTI and fMRI techniques in diagnostic practice is connected to limitations inherent in the technology: its methodological reliability and validity.

On the whole, the analysed verdicts display diverging views about which post-imaging data processing method provides the "best repeatability and clinical reliability" (case 1) for DTI examinations. In one of the verdicts, a court-appointed expert X (a neuroradiologist) holds DTI to be "an experimental technology, its examination methodology and certitude of findings have not

been validated. [Expert X] states that it is unclear which computational methods are reliable and how results from different facilities could be compared” (Case 2). Further along, the judge explicitly notes the expert stating that independent validation of DTI findings is not possible, because the resulting DTI images can only be examined in the very same apparatus and using the same imaging software version that delivered the initial findings. In judges’ accounts, experts both criticize and endorse drawing conclusions about the capacities of individual patients/plaintiffs based on DTI examination. Thus, “the reliability of DTI analysis depends decisively on the specific examination method employed by the facility/unit that performs the analysis; [it depends] on the repeatability of the method and size and comprehensiveness of the [control group], with which individual results are then compared” (case 9).

Most if not all positive evaluations concerning the possibility of making individual diagnoses with DTI are met by negative opinions. For example, some experts claim that control groups are in reality too small and too varied in their clinical characteristics to allow for individual diagnostic conclusions (e.g. case 1). Medical scientists and/or practitioners are locked in a controversy about what levels of measurement values constitute a significant enough deviation from normal levels to warrant an indication of brain injury in patients or plaintiffs (e.g. cases 1, 3), and some experts deny the validity of singular findings in diagnosis altogether and claim that DTI can only be used in comparing groups (e.g. case 7).

In the analysed cases, etiological non-specificity of DTI findings is another major topic regarding the validity of the examination method. The lack of validity is borne out of questioning the relation between DTI findings and the theory that the plaintiff indeed suffers from TBI. Thus, judges take note the claims made by the defendants’ experts that DTI cannot determine the actual cause of the observed tissue damage: “Many normal states, such as age or physical and psychiatric illnesses cause equivalent changes in brain tissue structure” (case 9). Other alternative causes, such as pain and PTSD are also considered factors that influence DTI findings. (Case 5.) Therefore, DTI findings that indicate changes in brain tissue do not necessarily indicate an actual injury, let alone a specific cause of the findings.

In some case verdicts, plaintiffs’ experts counter the argument about non-specificity by referring to the fact that there are no single examination methods at all which alone would suffice as a basis for diagnosis. Also, and especially in the case of young and healthy individuals, alternative causes are not, in terms of common sense, probable alternative causes of DTI findings. This is also something that many featured medical experts agree about, namely that strongly deviating DTI findings in

conjunction with other neurological and neuropsychological examinations do have some diagnostic value after all (e.g. case 3).

All of the themes (no cross-validation of findings, applicability to making individual diagnosis unclear, etiological non-specificity), discussed here non-exhaustively, do have an effect on how judges value DTI findings as a proof of brain injuries. This is explicitly so in some of the verdicts: “the district court considers it to have been shown, that DTI examinations cannot be used in diagnosing individuals, however, in conjunction with proper neuropsychological examinations DTI findings should be considered meaningful. It is not possible, though, to determine what has caused the findings” (case 3).

The difficulties associated with repeatability of DTI results and validity of the examination method result from the technology being experimental and not validated or established in clinical use. This makes for unreliable expert claims: in many of the eleven analysed cases, supporting TBI (or more specifically DAI) diagnosis with DTI findings backfired, as the opposing parties’ experts credibly pointed to the suspect character of the technology. The result is a credibility failure (Kirkland, 2012) for the plaintiffs’ experts and their claims about the presence of injury. However, arriving to such conclusions based on contested evidence and testimonial requires that judges manage to enact a legitimate understanding of alternative interpretations of evidence. Based on my analysis, I contend that the TBI case *judges possess a discursive capacity or ability to process interpretative flexibility in expert evidence and testimonial*. This ability is something the predefined criteria approach (e.g. Tadei et al., 2016) is unable to recognize due to giving no attention to how *judges themselves flexibly interpret evidence and use evaluative criteria to attribute credibility to some claims and withhold it from some*.

The other three Daubert criteria – *peer review*, *error rate* and *scientific/practical acceptance* – all receive a similar handling from the TBI case judges. I will briefly go through these next. Issues related to the three criteria are subject to alternative interpretations of evidence. Both parties’ experts support their side of the case by drawing on medical literature, as expected in the adversary setting of civil litigation in a Finnish district court, and in a dispute involving experimental technology. The lack of scholarly consensus on advanced MRI technology, or at least divergence in the interpretations that testifying experts make based on scientific studies, also marks considerable flexibility as a central feature of using *peer review* as an evaluative criterion.

The trust in both DTI and fMRI examination technologies also depends on them being free from systematic or recurring error. *Error sources and rates* are abundantly discussed in the TBI cases;

mostly they seem to be connected to errors in trained judgment, allegedly resulting from a lack of competence in the practice of diagnosing injury. Verdict discussion of experts' claims about possible error rates and sources displays judges' explicit reasoning about who and what to believe. Pointing to error sources can diminish the credibility of some claims and lead to a questioning of an expert's competence, but claims and knowledge about error sources can also signify individual experts' trained judgment and possibly mitigate opposing claims that seek to discredit the reliability of theories and examination methods such as DTI. Put differently, the case-specific significance of error rate is subject to interpretative flexibility on the part of the judge.

In the case of *consensus*, Tadei et al. (2016) differentiate between scientific and practical acceptance of a theory. Scientific and practical communities often have considerable overlap; in the TBI cases the featured experts regularly come from top university hospitals or other comparable institutions, and top-level clinical work is connected to making medical research. The relevant questions for judges are: what makes a community, who has epistemic authority to speak for it, and what does the community see as a proper perception about a given theory or technology? (cf. Jasanoff, 2015: 1728). The importance of consensus claims is reflected in the analysed verdicts' abundant discussion about whether the discussed medical knowledge and practices are generally accepted in the respective communities. With regard to the DTI, the judges' discussion in the eleven analysed verdicts indicates that there is at least a provisional consensus about the questionable reliability of the imaging technology based on its experimental developing character, while the discussion on fMRI displays less coherence in interpretations. In the eleven verdicts, the TBI court judges display a good control and grasp of this interpretative flexibility in determining the relative credibility of expert claims.

Results and conclusions

I set out by stating that different approaches to analysing judges' practice of evaluating expert evidence and testimonial produce differing understandings about how and how well judges perform evaluation. This, in turn, has consequences for perceptions of how judges should be educated and trained regarding expertise in court. To substantiate these claims, I contrasted two accounts discussing judges' evaluation of expertise: first, Tadei et al. (2016) exemplified the use of predefined analytical categorization into which judges' performance is then compared. I also showed that this approach is characteristic of a number of earlier studies that examine expertise in the courtroom context. Second, I provided an actor-based account of TBI cases in the Helsinki

district court, which described the way in which judges flexibly interpret evaluative criteria in discussing, making sense and ordering expert claims in arguing for their decisions.

Tadei et al. (2016) concluded that judges fare poorly at evaluating scientific evidence and testimonial, and do not understand the required scientific-analytical principles. The criticism I levelled at the study's analytical premises led me to conclude that contrary to the authors' suggestions, first, judges probably derive less from the technical training they receive than the authors hope for. The discussed reason was that to gain such a level of insight as to be able to evaluate the veracity of expert claims requires training and experience of working in the relevant professional community, and this kind of knowledge cannot be acquired through textbooks and short training courses. Secondly, there is no point in institutionalizing evaluation criteria (such as the Daubert criteria) to Finnish CJP. General evaluation criteria do not help in evaluating the veracity of knowledge claims, as also shown by the results of Tadei et al. (2016), because expertise is always specific to a field (or field of application), and not something general captured by abstract principles (Lynch, 2014). I concluded that teaching judges to use such criteria can lead to the criteria functioning as a legitimating and cloaking rather than a revelatory mechanism. This criticism applies more broadly to studies with a similar normative analytical setup that call for improving judges' (or jurors') scientific literacy (e.g. Chin, 2014; Gatowski et al., 2001; Wingate-Thornton, 2004). The criticism is also relevant cross-culturally, across different types of legal systems and jurisdictions (cf. Haack, 2005; Jasanoff, 2005).

The actor-based approach provides an account of how the examined actors (judges) themselves define and use analytical categories and criteria in evaluating expert evidence. My examination of the DAI and DTI/fMRI-related discussion in the Helsinki district court TBI case verdicts reveals that scientific concepts are discussed by judges at length, and specifically discussed as factors that influence the ordering of the presented evidence, and subsequently also the case resolution. *The empirical analysis shows that judges evaluate expertise by 1) flexibly interpreting evaluative criteria and 2) taking stock of interpretative flexibility in expert evidence and testimonial. It is the discursive management of the two that characterizes judges' evaluation practice.*

Judges need to work this flexibility into discussing expert claims, because they most often lack a deep understanding of the featured expertise and its field of application. Collins (1981: 4) states that the alternative interpretations of evidence that typify controversies reveal the negotiated, essentially cultural boundaries of what constitutes *scientific legitimacy*. The proof value of DTI, understood as an experimental and controversial technology in the field of TBI diagnosis and care, clearly depends on in-court negotiation of DTI evidence that determines the credibility of DTI findings.

Last edit of accepted manuscript. Published in Public Understanding of Science 28(8) 2019. Do not cite.

Re-phrasing Collins, an account of how judges discuss alternative interpretations of evidence also reveals the case-specific as well as culturally and institutionally bounded *evaluation practices* of trial judges; the specific professional manner of producing credibility and legitimacy in the process of reaching closure and arguing for the verdict (cf. Bal, 2005).

Despite my criticism, I do not claim that Tadei et al. (2016) and similar mentioned studies have no value as studies of judges' evaluation of scientific testimonial and evidence; courts' scientific literacy does matter a great deal, and in many disputes both inside and outside of court scientific evidence should carry considerable weight in relation to other sorts of evidence. There is a rationale, at least on the level of public intention, to inviting expert witnesses to court. More accurate knowledge and improved understanding of the disputed issue not only adds credibility to the claims of parties in a dispute, but also helps the fact finder to edge closer to correct understanding of the issue.

However, my analysis identified the unwanted consequence that might result from the predefined criteria approach, that is, the criteria becoming a legitimating mechanism that merely masks the original meta-expert problem of how to proficiently evaluate difficult expert evidence and testimonial. Therefore, prescriptions that call for an improvement in judges' scientific literacy should be informed by studies on how judges actually engage expertise in specific case types, and by case-specific discussions of interpretative flexibility. The realization that there is more to the courts' evaluation of expertise than a literacy deficit, should correct for the reductive understanding that results from normative approaches employing predefined analytical criteria.

My purpose here has been to show, first, how different approaches to analysing legal courts' evaluation of expertise are productive in the sense that they generate influential understandings of the issue, and second, that interpretative flexibility is important in understanding judges' evaluation practice and the role of scientific expertise for courts' decision making. Further studies should address the manner in which the flexibility of evidence *and* evaluative criteria is limited (cf. Collins, 1981; Baker, 2011) in court in order to allow for closure in a case. The conceptual discussion and results of this paper are also relevant to comparable investigations in other contexts; for example, in journalism and education non-experts encounter similar problems of assessing and using expertise, even if the respective goals, constraints and practices of the newsroom, the classroom and the courtroom would be different.

References

Helsinki district court: eleven traffic insurance compensation verdicts from 2014-2017 Diary numbers for eleven Helsinki district court traffic insurance compensation verdicts from 2014- 2017. All listed verdicts, all in Finnish, can be acquired from the Helsinki district court registry: 14/54740 18.12.2014 L 11/32949; 15/52847 4.12.2015 L 11/32937; 15/55033 16.12.2015 L 11/32961; 15/45215 26.10.2015 L 11/48585; 16/21038 9.5.2016 L 11/32945; 16/29950 17.6.2016 L 12/53220; 16/55891 14.11.2016 L 14/49517; 16/8354 26.2.2016 L 11/32940; Concept summary: 15.3.2017 20.2.2014 L 13/51074; Provisional verdict: 17/12756 9.3.2017 L 14/23057; 17/314 5.1.2017 L 11/32952

Bal R (2005) How to Kill with a Ballpoint: Credibility in Dutch Forensic Science. *Science Technology and Human Values* 30(1): 52-75.

Baker JD (2011) Tradition and Toxicity: Evidential Cultures in the Kava Safety Debate. *Social Studies of Science* 41(3): 361–384.

Bartlett A., Lewis J, Reyes-Galindo L and Stephens N (2018) The locus of legitimate interpretation in Big Data sciences: Lessons for computational social science from -omic biology and high-energy physics. *Big Data & Society* 5(1). <https://doi.org/10.1177/2053951718768831>

Bauer MW, Allum N and Miller S (2007) What can we learn from 25 years of PUS survey research? Liberating and expanding the agenda. *Public Understanding of Science* 16(1): 79-95.

Chin J (2014) Psychological Science's Replicability Crisis and What It Means for Science in the Courtroom. *Psychology, Public Policy and Law* 20(3): 225-238.

Cole S (1992) *Making Science: Between Nature and Society*. Cambridge, Mass.: Harvard University Press.

Cole SA and Bertenthal A (2017) Science, Technology, Society, and Law. *Annual Review of Law and Social Science* 13: 351-371.

Collins HM (1981) Stages in the Empirical Programme of Relativism. *Social Studies of Science* 11: 3-10.

Collins HM (1985) *Changing Order. Replication and Induction in Scientific Practice*. London: Sage, 1985.

Collins H (2011) Language and practice. *Social Studies of Science* 41(2): 271-300.

Collins HM and Evans R (2007) *Rethinking Expertise*. Chicago: University of Chicago Press.

Last edit of accepted manuscript. Published in Public Understanding of Science 28(8) 2019. Do not cite.

Collins HM and Evans R (2014) Actor and analyst: A response to Coopmans and Button. *Social Studies of Science* 44(5): 786-792.

Collins H, Evans R and Gorman M (2007) Trading zones and interactional expertise. *Studies in History and Philosophy of Science* 38: 657–666.

Daston L and Galison P (2008) *Objectivity*. New York: Zone Books.

Edmond G (2000) Judicial Representations of Scientific Evidence. *Modern Law Review* 63(2): 216-251.

Edmond G (2002): Legal Engineering: Contested Representations of Law, Science (and Non-Science) and Society. *Social Studies of Science* 32(3): 371–412.

Edmond G and Mercer D (1997) Scientific Literacy and the Jury: Reconsidering Jury ‘Competence’. *Public Understanding of Science* 6: 329-57.

Fleck L (1935) Scientific Observation and Perception in General. In Cohen R.S. and Schnelle T. (eds.), *Cognition and Fact – Materials on Ludwig Fleck*, 59-78. D. Reidel Publishing Company, 1986, pp. 59-78.

Gatowski SI, Dobbin SA, Richardson JT, Ginsburg GP, Merlino ML and Dahir V (2001) Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post-Daubert World. *Law and Human Behavior* 25(5).

Goldman A (2001) Experts: Which Ones Should You Trust? *Philosophy and Phenomenological Research* 63: 85–110.

Haack S (2005) Trial and Error: The Supreme Court’s Philosophy of Science. *American Journal of Public Health* 95(S1): S66-S73.

Israel-Jost V and Kinzel K (2014) Normative Engagement and Descriptive Accuracy in Science Studies. In Soler L, Zwart SD, Lynch M and Israel-Jost V (eds.) *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science*. New York: Routledge, pp. 114-122.

Jasanoff S (2005) Law’s Knowledge: Science for Justice in Legal Settings. *American Journal of Public Health* 95(S1): S49-S58.

Jasanoff S (2015) Serviceable Truths: Science for Action in Law and Policy. *Texas Law Review* 93: 1723-1749.

Jones OD, Marois R, Farah MJ and Greely HT (2013) Law and Neuroscience. *Journal of Neuroscience* 33(45): 17624-17630.

Last edit of accepted manuscript. Published in Public Understanding of Science 28(8) 2019. Do not cite.

Kirkland A (2012) Credibility Battles in the Autism Litigation. *Social Studies of Science* 42(2): 237–261.

Kovera MB and McAuliff BD (2000) The Effects of Peer Review and Evidence Quality on Judge Evaluations of Psychological Science: Are Judges Effective Gatekeepers? *Journal of Applied Psychology* 85(4): 574-86.

Kritzer HM (2007) Toward a Theorization of Craft. *Social & Legal Studies* 16(3): 321-340.

Law J (2008) On Sociology and STS. *The Sociological Review*, 56(4): 623–649.

Loiva J (2012) Asiantuntijatodistelusta De Lege Ferenda. [On expert evidence de lege ferenda.] *Defensor Legis* 5/2012: 639-665.

Lynch M (2014) From Normative to Descriptive and Back. In Soler L, Zwart SD, Lynch M and Israel-Jost V (eds.) *Science After the Practice Turn in the Philosophy, History, and Social Studies of Science*. New York: Routledge, pp. 93-113.

OMML 38/2015: *Tuomareiden koulutusjärjestelmän kehittäminen*. Oikeusministeriön julkaisu 38/2015. ([Developing judges' educational system]. Publication by Finnish Ministry of Justice 38/2015] <http://urn.fi/URN:ISBN:978-952-259-465-5>

Rask R (2011) Asiantuntijatodistelun Arviointi. [Evaluating Expert Testimonial.] In Lahti R and Siro J (eds.) *Asiantuntemustieto ja Asiantuntijat Oikeudessa*. Helsinki: Helsingin hovioikeus [Helsinki Appellate court], pp. 11-35.

Tadei A, Finnälä K, Reite A, Antfolk J and Santtila P (2016) Judges' Capacity to Evaluate Psychological and Psychiatric Expert Testimony. *Nordic Psychology* 68(3): 204-217.

Taipale J (2019): Judges' socio-technical review of contested expertise. *Social Studies of Science* 49(3): 310-332.

Turner S (2001) What is the Problem with Experts? *Social Studies of Science* 31(1): 123-149.

Uusitalo K (2013) Näytönarviointi ja näyttökynnys lapsiin kohdistuneissa väkivalta- ja seksuaalirikoksissa. [Evaluation of evidence and burden of proof in violent and sexual offences towards children.] In Mikkola T, Konttinen E, Uusitalo K, Väisänen T, Mattila T, Nieminen A and Puro K (eds.) *Lapsen Asema Kansainvälistyvässä Maailmassa*. Helsinki: Helsingin hovioikeus [Helsinki appellate court].

Last edit of accepted manuscript. Published in Public Understanding of Science 28(8) 2019. Do not cite.

Väisänen T and Korkman J (2014) Eräitä Todistajan Kertomuksen Arviointiin Liittyviä Kipukohtia Oikeuspsykologisen Tiedon Valossa. [Select difficulties in evaluation of witness testimonial in the light of forensic psychology knowledge] *Defensor Legis* 5/2014: 721-739.

Wingate PH and Thornton GC (2004) Industrial/Organizational Psychology and the Federal Judiciary: Expert Witness Testimony and the Daubert Standards. *Law and Human Behavior* 28(1): 97-114.

Wortzel HS, Tsiouris J and Filippi CG (2014) The Potential for Medicolegal Abuse: Diffusion Tensor Imaging in Traumatic Brain Injury. *American Journal of Bioethics (AJOB) Neuroscience* 5(2): 9-15.