

Instr Sci (2013) 41:989–1007
DOI 10.1007/s11251-013-9266-5

Development of legal expertise

Andreas Glöckner · Emanuel Towfigh · Christian Traxler

Received: 12 April 2011 / Accepted: 3 January 2013 / Published online: 12 January 2013
© The Author(s) 2013. This article is published with open access at Springerlink.com

Abstract In a comprehensive empirical investigation ($N = 71,405$) we analyzed the development of legal expertise in a critical 1-year period of academic legal training in which advanced law students start practicing to solve complex cases. We were particularly interested in the functional form of the learning curve and inter-individual differences in learning. Performance increases monotonically with the number of practice exams following a slightly concave learning curve without any considerable kinks. Considering the performance development over time, however, the curve is not monotonic and shows intermediate drops in performance. We provide evidence which suggests that these drops are due to cyclic drops in motivation. There are about equally sized marginal effects of practicing law exams in general and practicing exams in the specific area of law. However, students with high (vs. low) initial performance profit more from practicing exams within a specific area of law whereas students with low initial performance profit more from practicing exams in general. The concave increase in performance with the number of practicing exams is mainly driven by individuals with low initial performance. Those with high initial performance mainly display a linear learning trend. We discuss the practical implications of these findings for academic legal training.

Keywords Expertise · Legal decision making · Training · Learning curves

A. Glöckner
Georg-Elias-Müller-Institute for Psychology, Georg-August-University Göttingen, Gosslerstrasse 14,
37073 Göttingen, Germany

A. Glöckner (✉) · E. Towfigh · C. Traxler
Max Planck Institute for Research on Collective Goods, Kurt Schumacher Str. 10, 53113 Bonn,
Germany
e-mail: gloeckner@coll.mpg.de

C. Traxler
Department of Economics, Philipps-University Marburg, Am Plan 2, 35032 Marburg, Germany

Introduction

When solving complex cases, legal experts such as judges, attorneys and state prosecutors face a challenging task. First, they have to take into account multiple pieces of information about the case that are usually contradictory, incomplete, and made available in multiple formats. Second, even in highly codified areas of law it is often not obvious which legal rules should be applied and how they should be interpreted. Third, after solving these issues of fact-finding and legal evaluation, legal experts have to be able to form convincing arguments based on their interpretation of the case and the goals they aim to achieve. Consequently, legal expertise does not only mean knowing the law but it also includes specific skills for analyzing cases and to effectively write convincing legal arguments. Already in their academic education—which in Germany takes about 7 years including a mandatory legal clerkship—law students can be assumed to acquire a large stock of tacit knowledge (Marchant and Robinson 1999) and to develop complex cognitive structures (cf. Ausubel 1963) that enable them to evaluate and solve complex cases.

Considering the high relevance of legal judgments for society and the large number of practicing experts in law, it comes as a surprise that the development of legal expertise is relatively under-explored. There are a few theoretical surveys that cover issues of legal expertise at least partially (Blasi 1995; Glöckner and Ebert 2011; Herbig and Glöckner 2009; Marchant and Robinson 1999; Pennington and Hastie 1993; Robbennolt et al. 2010; Spellman 2010). Experiments that shed light on the development of legal expertise are even rarer. We traced only three empirical papers that investigate the development of legal expertise and its influence on performance in solving complex legal cases (Marchant et al. 1991, 1993; Nievelstein et al. 2010; for details see below; for a further study not yet considered in this paper see Dickert et al. 2012). The scarcity of empirical research might be due to at least two complications: the recruitment of persons with varying levels of legal expertise is usually effortful and costly; in addition, a reliable outside criterion for performance in solving complex legal cases is hard to find.

In the current paper we take advantage of a particularity of the German system of legal education that allows investigating the development of legal expertise over time for single persons. Instead of applying a fine-grained analysis of changes in cognitive structures (e.g., Ifenthaler et al. 2009), we use a (highly powered) regression approach. Specifically, we analyze data from more than 70,000 exams written by almost 3,000 advanced law students at the Faculty of Law at the University of Münster. We investigate the development of legal knowledge and expertise (a) over time and (b) by the number of repetitions in solving legal cases. We use a comprehensive panel-regression approach to analyze the influence of both factors—time and the number of past exams—on the achievements in exams that test performance in solving comprehensive legal cases repeatedly. To learn more about the factors influencing these developments, we also investigate the influence of local weather conditions on performance, which should be related to motivational factors.

We start by briefly summarizing findings from previous studies and discuss potential functional forms of the development of performance. We then provide background information on the structure of legal training in Germany and the exams that are used in our analysis. We report results from our analysis and conclude with a brief discussion.

Previous findings and hypotheses

Legal expertise and performance

Expertise has been investigated in various domains (e.g., chess, Chase and Simon 1973; medicine, Patel et al. 1994; Schmidt and Boshuizen 1993; psychotherapists, Witteman and van den Bercken 2007). In the field of *Judgment and Decision Making*, many studies demonstrate the prevalence of judgment errors also in professional decision makers in law (e.g., English 2008; English and Soder 2009; Guthrie et al. 2000, 2007; Schweizer 2005). Nevertheless, only a few studies provide a more detailed view on the development of legal expertise. In a series of experiments on applying tax law (Marchant et al. 1991, 1993), it was shown that experts have a distinctive advantage over novices in applying a potential source for analogy to new cases. However, legal expertise also seemed to have a downside in the sense that in some situations performance declines were observed for experts with a great deal of experience. This was explained by the fact that the experts had proceduralized the rule they applied in these situations (Anderson 1987) and did not look with sufficient care for alternatives. Hence—similar to findings in other domains (Frensch and Sternberg 1989)—increased experience can also have disadvantageous effects in variants of situations in which behavior deviating from often repeated standards or routines becomes necessary (Betsch et al. 2001; see also Einstellung Effect, Bilalic et al. 2008a).

A recent study investigating the role of conceptual knowledge shows that experts solve legal cases better than novices (Nivelstein et al. 2010). In ‘unsupported’ circumstances, in which participants had to rely on their knowledge only, novices and advanced students performed worse than domain experts, although even experts’ performance was rather low. Interestingly, however, in this unsupported condition advanced students (i.e., 3rd year law students) tended to perform worse than novices, indicating a discontinuity in the performance development. This difference was not significant ($t(20) = 1.05$, $p = 0.15$, one-sided; own re-analysis), which might, however, also be attributed to a lack of power in the analysis. The possibility to use additional information sources (e.g., law books) in a second condition positively affected performance only for advanced students, but not for novices. This might indicate that conceptual knowledge is a prerequisite for effective use of such additional sources.

Research questions

Given these results, it can be assumed that the ability to solve legal cases (i.e., performance) improves with legal training. However, some questions remain:

- (a) Does legal expertise increase monotonically with training and—more generally—what is the functional form of this development curve?
- (b) Are there differences in the learning curve if we consider the evolution of legal expertise over the number of practice trials versus the general time for studying?
- (c) How much do law students learn from area-of-law specific case-solving practice and how much from general case-solving practice? Are there inter-individual differences?

All three questions—and particularly the second and the third one—have not been sufficiently addressed in previous research. The identification of more or less important drivers of learning as well as of divergences between general studying time and practice experience could help optimizing the mix of time spent on general learning and the number of realistic practice trials within a specific area of law or in general. Theoretically,

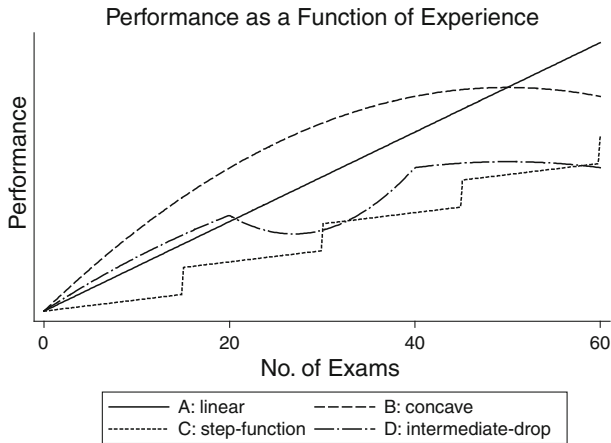


Fig. 1 Possible functional forms of the relation between performance and experience

divergences could help to identify drivers of performance (i.e., mere practice repetition within or above areas of law vs. learning time) and discontinuities (e.g., performance drops).

In the current study, we investigate these questions by following the performance of 4th-year to 6th-year law students over a year-long period in which they practiced solving legal cases. Note that students were not yet real experts by standard definitions of expertise (which often require 10 years of studies and/or experience, Ericsson and Smith 1991; Simon and Chase 1973), but we were able to trace their development in a critical period towards becoming experts.

Hypotheses

There exist different theoretical assumptions concerning the functional form describing the effect of experience on performance in the development of expertise and previous findings are partially inconclusive (details see below). As indicated by the results of Nievelstein et al. (2010) and findings in many other domains (e.g., Bilalic et al. 2008a, b; Schmidt and Boshuizen 1993; Wittman and van den Bercken 2007), performance does not always increase monotonically over time.

Nevertheless, one reasonable hypothesis is the case where performance increases linearly with experience (Fig. 1A). Such a development has been found, for example, concerning the influence of experience on the accuracy of medical diagnoses (Schmidt and Boshuizen 1993). Alternatively, performance development could follow a concave learning curve, with larger increases in the beginning and smaller improvements up to a certain value at a later stage (Fig. 1B). Such a functional form has been observed, for example, for the influence of total hours of serious study on current chess skill (i.e., a logarithmic relation, Charness et al. 2005) and is an often assumed functional form for expertise development (see Ericsson et al. 1993) as well as for skill acquisition (e.g., Newell and Rosenbloom 1981, in the form of a power-law). Alternatively, performance might even decrease from a certain point on, as shown in work on the relation between experience and job performance particularly for low complexity jobs (Sturman 2003). Discontinuities in the function in which performance sharply increases in a kind of step-function (Fig. 1C)—as shown, for example, in animal learning (Gallistel et al. 2004)—represent a further possible performance pattern. Another interesting case are

temporary drops in performance (Fig. 1D)—as they have been found, for example, for diagnostic performance of psychotherapists (Witteman and van den Bercken 2007) and in (temporarily increased) response times for perceptual inference tasks (Choi 1993). Such discontinuities can, for instance, be due to entering and leaving qualitatively different stages in the development of knowledge structures/cognitive structures (e.g., Baylor 2001) or motivational effects (e.g., Sturman 2003).

Due to the lack of clear predictions concerning the functional form, we conducted informal interviews with advanced law students and discussed the issue with experienced legal scholars in Germany. Based on their subjective estimation, we started with the hypothesis that there should be a general increase in performance with a temporary intermediate-drop (D).

We are not aware of any study investigating whether only area-specific or also general practice contributes to the development of legal expertise, and if so by how much. We start with the hypothesis that both kinds of practice make unique contributions to performance. Concerning inter-individual differences, we are particularly interested in differences in learning between students with initially high versus medium versus low exam performance. We expect differences in that students with good initial performance might benefit more from practice.

Background information on legal education

The system of German legal education consists of two steps. First, students enter law school for theoretic and doctrinal training in law. Time in law school encompasses a period of 4–6 years, culminating in the “First State Exam in Law” (*Erste juristische Staatsprüfung*), organized by each state’s Ministry of Justice and taken by the Appellate Courts.¹ After the first state exam, successful candidates may enter a 2-year period of practical training in a phase of government-paid “Legal Clerkship” (*Rechtsreferendariat*) with stages at a court, an office of the public prosecutor, in public administration and with a law firm. At the end of this legal clerkship stands the second state exam, again taken by the Appellate Courts. Only after successful completion of this exam is one considered a fully fledged legal practitioner jurist (*Volljurist*) who can fill any position as a judge, public prosecutor, and attorney or as a lawyer in public administration. The whole legal education, including exam periods, takes at least 7 years.

The results of the state exams are by and large the only signals considered on the job market. While credits collected by taking university exams are a prerequisite for registering for the first state exam, they do not enter into the results of the state exam and thus only play a minor role in helping students assess their performance. The state exam can be taken twice. After two failures, the entire legal studies end without any degree and are thus rendered worthless. Therefore, preparing for the state exam plays an overwhelmingly important role. Preparation time for the first state exam is not limited, though usually it will take about 1 year.²

¹ The exam consists of five to eight written exams of 5 h each (the number depending on the state), and an oral examination of approximately 90 min per candidate.

² This is due to the fact that if the application to take the first state exam is registered before the end of the 4th year of studies one is awarded a “free trial” (*Freiversuch*), meaning that a failure will not be counted; successful candidates may take the exam again if they are not satisfied with the results, an option not available when not taking the free trial.

It is in this context that universities (and a number of private coaching services) help students prepare for their first state exam. One of the training possibilities offered by many universities to their students throughout the year, including vacation times, is to take unpublished exams from earlier state exams provided by the Appellate Courts in a “Written Examination Course” (*Klausurenkurs*). The course is open to all students who have completed the university’s legal training and thus meet the requirements for registering their application for the first exam. Parallel to taking these exams, students pursue their general studies in preparing for the state exam; their preparation, however, is not related to the somewhat arbitrary content of the exam that may come to them as a surprise the same way as it does during the real exam.

The University of Münster, Germany’s largest faculty of law, provided us with the data for our study. In Münster, the preparation exams are offered in what can be considered a real-life setting: working time is limited to 5 h, there is supervisory staff, and only the legal texts admissible in the first state exam may be brought, while further text books etc. are not allowed. The exams are then corrected within 2 weeks according to the sample solutions provided by the Appellate Court with each exam; then the exams are returned to candidates. Corrections are carried out by senior teaching assistants, who have at least the first state exam, but who do not usually correct state exams; for each correction, students have to pay 5 Euros. In addition, the exam will be discussed in a special 2-hour class by an experienced first state exam examiner.

As in the state exam, these mock exams are offered in the three topical areas common in German law: civil, public and criminal law. Two exams are offered each week, one being from the area of civil law, the other alternating from the area of public or of criminal law. While attending the examination course is not mandatory, basically all law students take some of these exams to prepare properly for the final exam, both with a view to the contents and the difficulty level, but also in order to practice proper time and resources management. The university recommends taking around 20 training exams per topical area before entering the first state exam. Students may enter the course at any time during the year, depending on their level of preparation, and they may skip any of the exams without further notice or reason. Moreover, they may start taking the exams in one area of law at one period in time, and in another area later on.

In summary, it can be assumed that the exam grades reliably measure students’ performance development in a critical phase of their studies in which they are trained to apply legal rules to hypothetical, but realistic cases. The exams are sufficiently often and essentially randomly assigned to participants to allow for a reliable analysis of performance development using panel data methods.

Data

The University of Münster provided us with anonymized data on all exams that were held between October 1999 and January 2008. In total, we observed roughly 80,000 graded exams. As we were interested in the development of performance over time, we focus on a sample of students who took at least 10 exams and at most 62 exams (i.e., we deleted the top percentile and the lowest tercile of a highly skewed distribution). This left us with a qualified sample of 71,405 exams from 2,979 students. The students in our sample took on average 24 exams with an average time between the first and the last exam of 43 weeks. Most exams were taken in civil law. Sample descriptives are summarized in Table 1.

Table 1 Data sample description

Variable	Statistics
Exams (total considered set)	71,405
Students (percent in parentheses)	2,979
Female	1,508 (50.6 %)
Male	1,471 (49.4 %)
Exams per student M (SD) [MD]	24.0 (11.0) [22]
Duration	
Weeks between first and last exam M (SD) [MD]	43.3 (30.2) [37]
Areas of Law (percent in parentheses)	
Civil Law	30,226 (42.3 %)
Public Law	18,754 (26.3 %)
Criminal Law	22,425 (31.4 %)

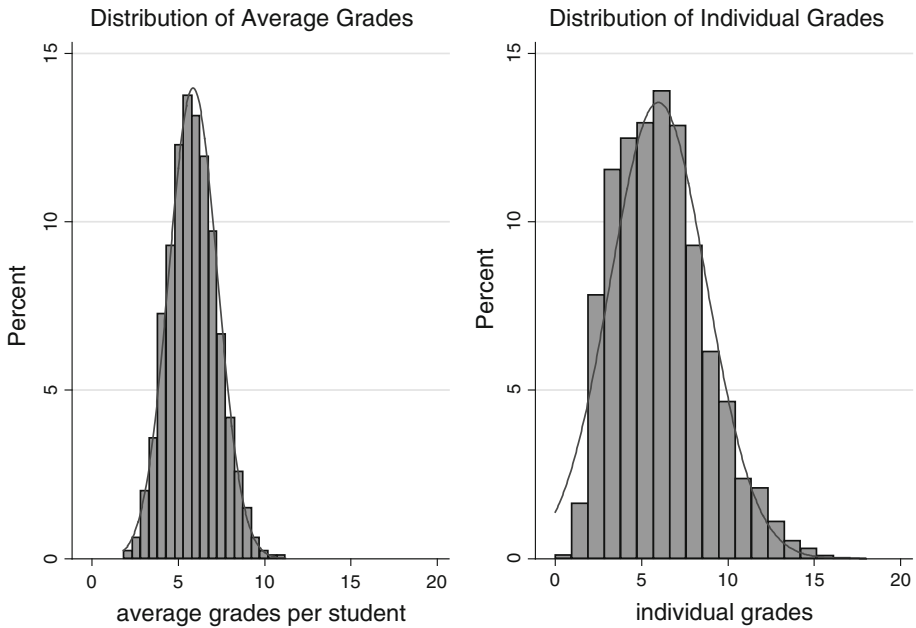


Fig. 2 Average and individual grades

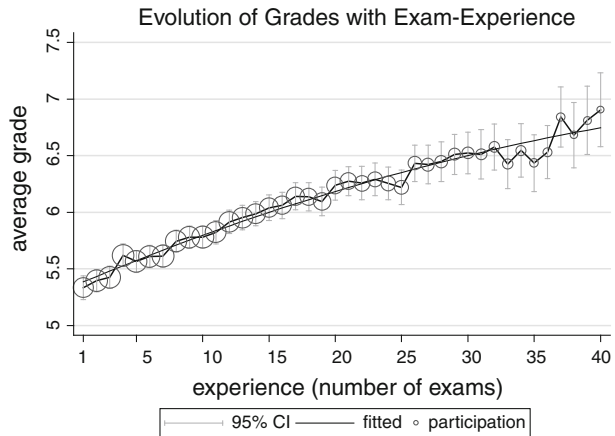
Performance in the exams is measured in grade points ranging from 0 to 18, with higher numbers indicating better performance. The left panel of Fig. 2 displays the distribution of the average grades of the 2,979 students in our sample (i.e., the average over all graded exams for each student). The figure shows a nearly normal distribution with a mean grade of 5.81. The right panel depicts the distribution of the 71,405 single exams. This distribution is slightly more skewed, but has a similar mean of 5.96 (median and modus corresponds to a grade of 6). More descriptive data on grades are summarized in Table 2.

Table 2 Summary of exam grades

Grades	Statistics
Overall M (SD) [MD]	5.96 (2.79) [6]
Grades by Area of Law (SE in parentheses)	
M _{Civil Law}	6.07 (0.03)
M _{Public Law}	6.16 (0.03)
M _{Criminal Law}	5.66 (0.03)
Grades by Sex (SE in parentheses)	
M _{Female}	5.83 (0.04)
M _{Male}	6.10 (0.04)
Grades by Exam Experience (incl. all areas of law)	
M _{Exams 1–10}	5.59 (0.03)
M _{Exams 11–20}	6.02 (0.03)
M _{Exams 21–30}	6.34 (0.04)
M _{Exams 31–}	6.62 (0.08)
Grades by Studying Time since first exam (incl. all areas of law)	
M _{Weeks 1–10}	5.60 (0.03)
M _{Weeks 11–20}	6.01 (0.04)
M _{Weeks 21–30}	6.19 (0.04)
M _{Weeks 30–}	6.05 (0.04)

SEs are cluster corrected at the individual level (Rogers 1993)

Fig. 3 Evolution of grades with experience (number of exams); error bars indicate 95 % confidence intervals (based on pooled standard errors)



Analysis

Performance development by exam experience

We first consider the evolution of performance, measured by the grade, over experience, measured by the number of exams written by each student. The experience-performance pattern based on individually graded exams is illustrated in Fig. 3. The data reveal a steady increase in average grades. As students gain more experience (i.e., have written more exams), their grades increase almost monotonically. In their very first exam, students start with a grade that is significantly below 5.5. After 20 exams, they earn grades significantly above 6.0, and after 30 exams, the average grade reaches 6.5. There are some small kinks

along the improvement path; however, none of these is significant, in the sense that the average grade at the kink differs significantly from the grades in the previous or in the following exam.

Confronting Fig. 3 with the possible patterns displayed in Fig. 1, our data clearly reject a performance–experience pattern that follows a step-function or has intermediate drops (curves C and D in Fig. 1, respectively). The data rather point to a linear or a slightly concave pattern.

We further explore the performance–experience relationship in a regression analysis. As a dependent variable we use the grade of student i in each of her t th graded exams, which corresponds to the n th exam within a topical area a , during term y (for simplicity subscripts y are omitted in the notation). Our main focus is the effect of experience, the overall number of completed exams ($Exam_{it}$), on the student's grade achievements. We consider specifications including a linear and a quadratic term.³ In addition, we control for the total number of exams within each topical area of law ($ExamNo_{ian}$) and the number of participation in each specific exam ($Partic_{an}$). Furthermore, we account for season-specific effects and control for the weather conditions on the day of the exam (more precisely, for the average temperature, sunshine duration, rainfall, cloudiness and atmospheric pressure). The basic structure of our estimation equation reads

$$Grade_{itan} = \alpha + \beta_1 Exam_{it} + \beta_2 Exam_{it}^2 + \gamma ExamNo_{ian} + \theta Partic_{an} + \varepsilon_{itan} \quad (1)$$

Making use of the panel structure of our data (i.e., the fact that we observe a sequence of exams for each student), we additionally account for individual specific fixed effects, term-specific (winter vs. summer term), area-specific (civil vs. public vs. criminal law) and exam-sequence fixed effects (i.e., for the 1st, 2nd, etc. exam within each topical area within each term). The advantage of this large set of fixed effects is that observed and unobserved heterogeneity in characteristics (age, gender, ability, etc.)—driving time-invariant differences in average grades at the individual level—as well as unobserved exam-specific features (e.g., the difficulty of the 1st, 2nd, etc. exam within each topical area within each term)—that potentially affect the grades at one specific day—are absorbed by the large set of fixed effects and control variables.

Ordinary least square estimates of Eq. (1) (including the set of fixed effects discussed above) indicate a concave learning pattern (Table 3, specification 2). As compared to a linear learning model (Table 3, specification 1), the non-linear model does slightly better in explaining our data. The data further indicate different layers of improvement. On the one hand, there is a general improvement with any type of exam experience: having written 10 exams in, say, civil law helps students to earn better grades in their 11th exam, even if it is in, say, criminal law. On the other hand, there is an additional effect from area-specific learning: one more exam within a given area of law (e.g., civil law) makes students earn better grades at the next exam in this area. This effect occurs on top of the overall effect from general exam experience (Table 3, specification 3). Hence, there is support for the hypothesis that both area-of-law-specific and general exam practice contribute independently to performance development.

Performance development by time

The evidence from above gives a clear picture on grades being monotonically increasing in experience and that both area-specific and general exam practice contribute independently

³ We also included higher-order polynomials; however, these terms never turned out significant.

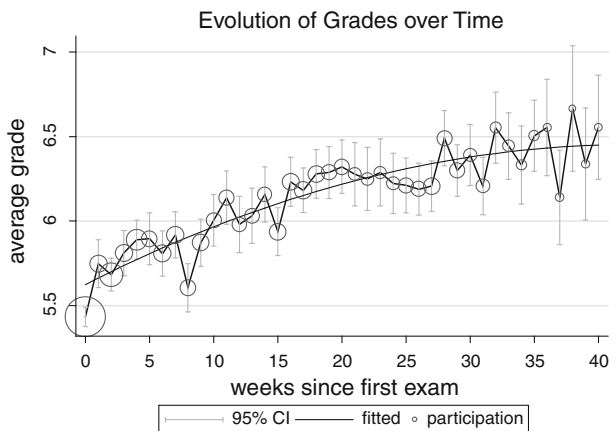
Table 3 Regression models estimating grades with five different specifications

Dependent variable: grade	(1)	(2)	(3)	(4)	(5)
Experience	0.027*** [0.002]	0.044*** [0.004]	0.034*** [0.004]	0.035*** [0.004]	0.035*** [0.004]
Experience-squared		-0.0004*** [0.000]	-0.0004*** [0.000]	-0.0004*** [0.000]	-0.0004*** [0.000]
Experience, area-specific			0.032*** [0.005]	0.031*** [0.005]	0.031*** [0.005]
Week 8				-0.238*** [0.068]	-0.973*** [0.222]
Week 8 × cloudiness (t - 1)					0.013** [0.004]
Cloudiness (t - 1)					0.0004 [0.001]
Area: public law	0.127*** [0.029]	0.120*** [0.029]	0.190*** [0.031]	0.191*** [0.031]	0.189*** [0.031]
Area: criminal law	-0.417*** [0.030]	-0.425*** [0.030]	-0.383*** [0.031]	-0.381*** [0.031]	-0.381*** [0.031]
Constant	6.657*** [1.211]	6.494*** [1.210]	6.516*** [1.210]	6.614*** [1.211]	6.729*** [1.222]
Observations	71,379	71,379	71,379	71,379	71,379
Adjusted R ²	0.2468	0.2472	0.2476	0.2478	0.2479

Ordinary-least-square regressions with robust standard errors (clustered at the student level) in squared brackets. Areas (public law and criminal law) are dummies with the reference area being civil law. All specifications include individual, term-specific, exam-specific and season-specific fixed effects (results not reported). Moreover, we control for the weather conditions on the day of the exam (average temperature, sunshine duration, rainfall, cloud amount as well as the atmospheric pressure; results not reported) and participation (number of participants for the exam; results not reported)

/ indicates significance at the 0.01/0.001-level, respectively

Fig. 4 Evolution of grades over time by week since first exam in the respective area-of-law; error bars indicate 95 % confidence interval (based on pooled standard errors)



to this development. However, following our second research question, we are also interested to identify whether there are effects of time taken for learning in an area of law on top of this experience effect (i.e., exam-solving practice; see also discussion of testing effects below). We thereby make use of the fact that not all students do take one exam per

week. Moreover, there may be public holidays interrupting the weekly exam rhythm. Figure 4 shows how grades evolve over time, with time being measured by weeks since a student's first exam within a given area of law. The figure confirms our previous result of a general, and slightly concave upward trend. However, the improvement path is now substantially less smooth than suggested by the pattern from Fig. 3. In particular, 8 weeks after the first exam in the area, there is a pronounced drop in average grades from 5.92 to 5.60.

To check this effect statistically, we extend Eq. (1) from above by including an indicator variable that accounts for a potentially specific effect on grades in the 8th week after the first exam in the respective area of law. Note that the regression accounts for many possible (e.g., exam-specific, student-specific, season-specific, term-specific, topic-specific) effects that could in principle drive this drop.⁴ The estimates confirm the impression obtained from Fig. 4: the stark decline in week 8 remains statistically significant and substantial in magnitude (Table 3, specification 4). Accounting for all other influences, we estimate an average drop of 0.24 grading points in week 8 after the first exam.

According to these estimates, the week-8-drop thus washes away the improvement that comes with an experience of 7.5 exams. This is also illustrated in Fig. 4, where the drop brings the average grade nearly back to the level observed for the very first exam. Hence, the data support the hypothesis that there are intermediate kinks in the performance development.

To double-check the stability of the 8-week drop we also analyzed the data using an alternative statistical approach, which is more common in educational research. Specifically, we analyzed specification 4 (Table 3) using a multi-level regression assuming a random constant and random slopes for experience and experience-in-area (and zero covariance between the three coefficients). We had no exact a priori hypothesis concerning the exact time of eventual performance drops. Hence, our selection of week 8 was data-driven, which might lead to an implicit accumulation of alpha errors. To account for this problem, we used a Bonferroni correction. Specifically, we corrected the alpha-error level for 40 tests (i.e., weeks 1–40; $\alpha_{corr} = 0.05/40 = 0.00125$) for identifying effects of drops by week (control dummies were also omitted). Even with this alternative way of analyzing the data, the 8-week drop turned out to be significant, $b = -0.35$, $z = -5.27$, $p < 0.00001$.

Potential reasons for the 8-week-drop

One could think of several different reasons that explain this drop at week 8: first, the drop might be due to knowledge restructuring; second, one might think of motivational effects; third, the observation might simply be driven by a selection effects: if good students realized they are tired after 8 weeks of learning and—optimizing their mid-run learning-output performance—decide to “take a break at week 8”, we will just have a change in the student composition in this week: with fewer smart students, average grades will fall.

⁴ One could think, for instance, of all civil law exams in the 8th week after the start of a new exam series to be substantially more difficult than other exams. However, students can (and do) start taking exams in this year-round preparation course at any time during the year, so this in itself makes such an explanation unlikely. Moreover, such effect would be statistically absorbed by the dummy for the 8th exam within the civil law series. Furthermore, one might suspect that there are some special regulations concerning exams after 2 months that might cause drops. Being a lawyer that has been involved in the administration of the examination, the second author of this paper knows both the formal rules and the informal common practices very well and can confirm that no such rules exist.

Given the fact that when beginning to practice exams, students start at quite different knowledge levels (and even in different years of their studies), the first explanation of a cognitive restructuring after 8 weeks seems somewhat unlikely. We can, however, not completely rule out this possibility given our data. Furthermore, additionally analyses (not reported) reveal that the drop concerns high, medium, and low initial performers and therefore is unlikely to be driven by high performers dropping out at week 8 (i.e., self-selection; see also participation rates in Fig. 4). We analyze the motivational hypothesis statistically in the following.

Motivational effects and weather To assess the importance of motivational issues for the week 8 drop, we exploit meteorological data from the city of Münster. We study to which extent the drop is more or less pronounced (or remains unaffected) when weather conditions were “good” on the days before the exam took place. To account for the fact that weather conditions might affect students’ mood, and thus their learning success and grades, we decided not to focus on sunshine or rainfall (which is known to affect peoples’ mood; see Keller et al. 2005). Instead, we measure the cloudiness on the day before the exam.⁵ Apart from the direct week 8 effect (captured by χ_1), our model also includes an interaction term that measures the specific effect of cloudy weather on the day before the exam in the 8th week (χ_2). The general impact of cloud coverage (beyond week 8) is measured by χ_3 .

$$\begin{aligned} \text{Grade}_{itn} = & \alpha + \beta_1 \text{Exam}_{it} + \beta_2 \text{Exam}_{it}^2 + \gamma \text{ExamNo}_{ian} + \theta \text{Partic}_{an} + \chi_1 \text{Week8}_{it} \\ & + \chi_2 \text{Week8}_{it} \times \text{Clouds}_{it}^{\text{prev.day}} + \chi_3 \text{Clouds}_{it}^{\text{prev.day}} + \mu \text{Weather}_{it} + \varepsilon_{itn} \end{aligned} \quad (2)$$

Ordinary least square estimates of Eq. (2) (including the same set of fixed effects as used before) yield a clear result (Table 3, specification 5): if the weather was good before the exam—i.e., if there was a day without any clouds on the sky—the drop in the grade was even more pronounced (0.97 grading points instead of 0.24, observed for the sample mean). The estimates further indicate that an increase in cloud coverage by one standard deviation on the day before the exam makes the week 8 drop shrink. Hence, if the weather during the last day of the studying phase is bad, students obtain better grades on next day’s exam. Stated differently: the better the weather was, the worse was the drop in week 8. It is important to note that the regressions do not indicate a general effect of the weather prior to the exam on the grade outcome.⁶ The effect only occurs before the week-8-drop.

The evidence therefore supports a motivational interpretation of the week 8 drop: it seems to be hard for Münster’s law students to focus on learning in the 8th week after they started with their first exam. However, if it is a gray day, it might be easier to stay in the library and focus, for instance, on criminal law for another hour, rather than chatting with fellow students at the *Aasee* (a beautiful lake in the heart of Münster).

⁵ Note that we obtain similar results if we consider the average cloud amount during the last 2 or 3 days (rather than 1 single day) before the exam. Recall further that we control for the weather conditions on the day of the exam (more precisely, average temperature, sunshine duration, rainfall, cloud amount as well as atmospheric pressure), which might affect the performance on that specific day. We also conducted the analysis including the interaction of cloudiness at the day of the exam and the 8 week dummy. This interaction did not turn out significant, indicating that the motivational effect seems to be limited to the preparation time for exams but does not concern carrying out the exam in itself.

⁶ One might be concerned that cloudiness is correlated with sunshine, which in turn might be correlated with students’ mood. If this were the case, one would expect a negative effect of a high level of cloudiness (and presumably bad mood) on the day before the exam to have a negative effect on grades. This is not the case: χ_3 is not significantly different from zero.

Other findings concerning the 8-week-drop

Assuming that the week 8 drop is driven by motivational aspects, it might be reasonable to suppose that it repeats in intervals of 8 weeks. We indeed find similar, but somewhat weaker drops for weeks 15 and 25 (see Fig. 4), though these drops did not turn out to be significant. This insignificance could be due to larger errors from aggregating heterogeneous individuals with overlapping motivational ‘cycles’ over multiple time periods. Looking at male and female students separately, we observe the week 8 drop for both of them. It tends to be somewhat weaker for female students, however. We also find the drop when considering the three topical areas of law—civil, criminal and public law—separately.

Drivers of learning performance

Effects of area-specific versus area-general practice

As indicated by the similarity of the coefficients for (general) experience with practicing exams and area-specific experience in Table 3, law students profit about equally from area-of-law specific practice and from general practice of exams (test for differences between coefficients: $p = 0.67$). Note, that coefficients indicate independent contributions of each factor to learning and furthermore note that in Germany exams in different areas of law have only little (if any) overlap in legal content since they concern completely different legal codes (i.e., different books). Therefore, it is quite remarkable that practicing exams outside the specific area of law has such a huge effect. The effect must be driven by developing domain general skills how to solve cases and/or an increasing general understanding of law.

Inter-individual differences

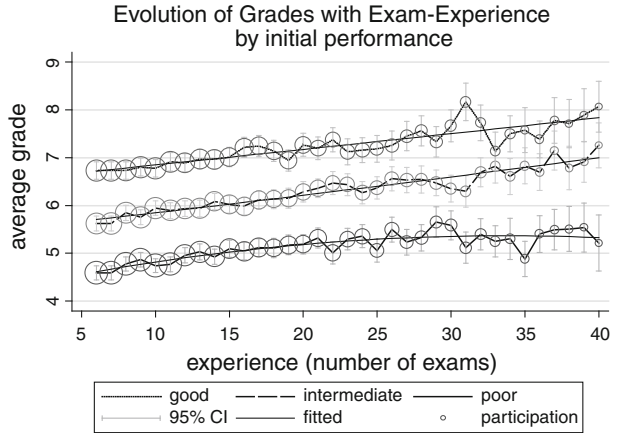
In a final step, we were interested in identifying inter-individual differences in learning and drivers of learning between different individuals. Knowledge concerning these differences could allow for individual-specific recommendations concerning efficient practice. Our analysis focuses on effects of initial performance. At the end of this subsection, we will briefly discuss gender effects.

To investigate the effects of initial performance, we split students in three equally sized groups according to their average scores over their first 5 exams. Students in the lower (average grade exams 1–5 <4.6), medium, and upper (average grade exams 1–5 >6.2) 33-percentiles were classified as low, medium, and high initial performers, respectively. This rough classification is mainly used for descriptive purposes. In the main statistical analysis, in contrast, initial performance is treated as continuous variable.

We first analyzed differences in learning between performance types. Figure 5 displays learning curves by exam experience (i.e., total number of exams taken). The different initial performance types are indicated by different lines. The graph shows that grades of good and intermediate initial performers increase stronger than grades of poor initial performers. Moreover, performance development for good and intermediate initial performers is more close to a linear trend whereas performance development for poor initial performers is concave. Hence, the overall slightly concave functional form (see Fig. 3) seems to be mainly driven by low initial performers.⁷

⁷ One might argue that the observed pattern is also driven by different performers dropping out the exam cycle at different points in time. While a duration analysis indeed indicates differences in the average

Fig. 5 Evolution of grades with exam experience for students with good, intermediate, and poor initial performance—capturing the top, mid, and lowest tercile of the distribution of average grades in the first five exams. Data from the first five exams that were used for classification are excluded. *Fitted lines* represent regression predictions including a linear and a quadratic term. 95 % confidence interval indicated based on pooled standard errors



Note, however, that the development by exam experience from Fig. 5 reflects the overall effect of exam experience, including the parts that the general exam experience effect shares with area-of-law-specific exam experience—both are naturally correlated. Hence, conclusions concerning differential effects drawn from Fig. 5 might be misleading. We therefore used a regression approach to investigate differential effects more thoroughly. Similar to the regressions reported in Table 3, grades are explained by the main effects of general exam experience (i.e., linear and squared function of number of exams) and area-specific experience (i.e., number of exams within a topical area of law). Most importantly, all three two-way interactions of these variables with the initial performance (measured by a continuous variable, the average grade obtained in the first five exams) are included in the model. Differential learning effects for students with different initial performance levels should be reflected in significant interaction terms. Beyond these interaction terms, the regression includes the same set of fixed effects and control variables as above. Further note that the inclusion of individual level fixed effects takes up the effect from the (non-interacted) level of initial performance. Estimation results are reported in Table 4.

On top of the positive linear effect of exam experience and the negative squared exam experience effect already established above, both factors interact significantly with initial performance. The positive interaction with the squared term indicates that the concave functional form is mainly driven by persons with lower initial performance and that the performance curve becomes more linear (and even convex) with increasing initial performance. The negative interaction with the linear effect indicates that the established effect of domain-unspecific experience due to taking exams outside the specific area-of-law is smaller for persons with higher as compared to lower initial performance. Stated differently (and in contrast to the impression conveyed by the overall effect in Fig. 5), those who initially perform poorly benefit relatively more from general practice than good initial performers.

Footnote 7 continued

number of exams taken between the different groups (medium (24.4) and high-initial performers (24.2) took on average one exam more than low performers (23.3)), these differences are quantitatively too small to drive the different learning curves.

Table 4 Regression models estimating grades with a specification for analyzing interindividual differences dependent on initial performance

Dependent variable: grade	(1)
Experience	0.0303*** [0.00561]
Experience-squared	-0.0003*** [0.00010]
Experience, area-specific	0.0312*** [0.00580]
Experience × initial-performance	-0.0094*** [0.00303]
Experience-squ. × initial-performance	0.0002*** [0.00006]
Experience, area-specific × initial-perf.	0.0056* [0.00320]
Observations	57,188
R-squared	0.289

Ordinary-least-square regressions with robust standard errors (clustered at the student level) in squared brackets. All specifications include individual-, term-, exam-, area- and season-specific fixed effects (results not reported). Initial-performance measures the average grade in an individual's first five exams. Each individuals' first five exams (used to compute the variable) are excluded from the estimation sample. Note further, that the (non-interacted) initial-performance effect is omitted, since it is absorbed in the individual level fixed effects. The regression also controls for the weather conditions on the day of the exam (average temperature, sunshine duration, rainfall, cloud amount as well as the atmospheric pressure; results not reported) and participation (number of participants for the exam; results not reported). All variables are centered

*/*** indicates significance at the 0.05/0.001-level, respectively

As expected, the area-of-law specific exam experience effect also replicates in this specification. More importantly, however, it interacts *positively* with initial performance. This means that those with a higher initial performance profit more from area-of-law specific practice as compared to initial poor performers.

Together with the previous finding, the regression analysis therefore indicates the following pattern: the weaker the performance during the first exams, the more strongly do the individuals benefit from general exam practice, and the less relevant is area-of-law specific practice for their improvement. Vice versa, for those with a higher initial performance general experience effects are less crucial whereas the benefits from area-of-law specific exams are more sizable. Hence, the performance development of the one group seems to more strongly benefit from general practice, the other group profits more from specific practicing.

Considering the role of gender, the data revealed a general gender effect, in that female ($M = 5.83$, $SE = 0.039$, $N = 1,508$) performed significantly worse than male ($M = 6.10$, $SE = 0.041$, $N = 1,471$) students, $F(1, 2978) = 21.17$, $p < 0.001$. Also both the general and the area-of-law specific learning effects of experience were higher for male students ($b_{unspec} = 0.0371$, $SE = 0.0058$; $b_{spec} = 0.036$, $SE = 0.0082$) as compared to females students ($b_{unspec} = 0.0312$, $SE = 0.0063$; $b_{spec} = 0.028$, $SE = 0.0072$).

Discussion

We investigated the development of legal expertise in a large sample of advanced law students over a period of about 1 year in which, aside from general studying, they practiced solving complex legal cases in written mock exams. We find that their performance increases monotonically with the number of mock exams they take. The performance-exam curve has a smooth, slightly concave shape without any kinks. When we consider the time since the first exam in the respective area of law, the development is much less smooth and shows a strong and significant drop at week 8: performance goes almost back to the starting level. Thereafter, performance quickly recovers to the pre-drop level. This effect seems to be driven by a motivational effect: it decreases if the weather is bad on the days before the exam, that is, when grey skies make the outside options to learning appear less attractive. However, we cannot rule out that the effect might also partially be driven by cognitive restructuring.

Our analysis further showed that area-of-law specific exam practice and general exam practice (i.e., doing exams in other areas of law) have about equally sized effects on the improvement in exam performance. This finding is important, since the contents of the different areas of law in Germany are quite distinct; they concern separate legal codes and involve completely different study books and commentaries. Hence, performance is not only driven by learning the specific content of the area of law but also by some more general kind of learning that could be due to general case-solving skills or general knowledge of law. To the best of our knowledge this is the first study demonstrating the contributions of such direct and more indirect learning effects in expertise development.

We also observe clear differential effects in that students with different levels of initial performance benefit differently from general exam practice and area-of-law specific exams. Students who initially perform better profited more from area-of-law specific exams, whereas grades for students with lower initial performance levels increased more by general exam practice. Moreover, the concave learning effect of practice on performance that is often discussed in previous studies was only observed for students with poor initial performance. Students with higher initial performance basically showed linear learning curves.

The study provides insight into an interesting part of the development of legal expertise, namely the time when students move from merely acquiring theoretical textbook knowledge to practicing the solving of legal cases. The large sample allows a fairly precise estimation of the learning curve and avoids pitfalls from the ad-hoc selection of measurement points. More importantly, however, the data allow for an in-depth analysis of the components driving performance development and inter-individual differences.

Besides allowing a more differentiated perspective on expertise development in general, and in the area of legal expertise in particular (an area that has been largely unexplored so far), the current findings also tie into work on testing effects (for a recent review see Roediger et al. 2011) by providing an analysis of the functional form of the effect of frequently applied tests on performance. Interestingly, the observed effects of repeated general and area-of-law-specific exams (in a sense ‘testing’) reported in Table 4 remain stable and significant even when controlling for the time since the first exam in the respective area of law, which is a proxy for (intensive) studying time involved in the respective area of law.⁸ Stated differently, the results concerning effects of exams

⁸ Specifically, most other coefficients remain stable when controlling for a linear and a squared term of weeks-since-the-first-exam in the respective area of law as well as their interactions with performance. Only the interaction of area-of-law-specific exams and initial performance is reduced substantially to non-significance.

presented in this paper also hold when (roughly) controlling for studying time. Hence, the results could be at least partially driven by testing effects on top of mere effects of studying the materials. The strong effect of area-of-law specific and area-general practice exams could, for instance, be plausibly explained by the fact that testing produces better organization of knowledge and that testing improves transfer of knowledge to new contexts (Roediger et al. 2011). The positive interaction of area-of-law specific practice with initial performance could furthermore be explained by the fact that knowledge structuring advantages might be particularly high for persons that start with a good basic knowledge in the respective area of law. Nevertheless, it is due to further research to investigate the relations between testing effects and expertise development in more detail.

Practical implications

Based on our results, we can provide practical and pragmatic advice for law students entering the practice exam phase. First, the money and time they invest in repeatedly solving mock exams is worthwhile. There is an (area-of-law) unspecific practice effect, in that each exam increases the performance score by roughly 0.5 %, although this amount decreases somewhat over time (particularly for persons with low initial performance). On top, there is an area-of-law specific repetition effect of also roughly 0.5 % per practice exam in this specific area of law. Hence, incremental improvements are relatively small. It should also be noted that these repetition specific improvement effects are predicted for students combining mock exam participation with “normal” studying (instead of taking exams only). However, as reported in the previous section, the results also hold when roughly controlling for intense studying time. Second, it seems advisable to take into account exhaustion and drops in motivation. Students might therefore consider scheduling time for somewhat more extended breaks in roughly 8-week intervals. Third, our results can be helpful for students’ general time planning. They could use their performance in early exams to predict the potential for improvements and the time they will need to realize their goals. Forth, students with good initial performance should focus more on area-of-law specific practice exams whereas students with poor initial performance seem to profit more from general exam practice in all areas of law.

Acknowledgments The authors are grateful to the Faculty of Law at the University of Münster for providing the extensive dataset that was made anonymous. Particular thanks go to Annette Barkey-Heine (Examination Office) and Ulrich Weber-Steinhaus (IT Department).

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Anderson, J. R. (1987). Skill acquisition: compilation of weak-method problem solutions. *Psychological Review*, *94*, 192–210.
- Ausubel, D. P. (1963). Cognitive structure and the facilitation of meaningful verbal-learning. *Journal of Teacher Education*, *14*, 217–222.
- Baylor, A. L. (2001). A U-shaped model for the development of intuition by level of expertise. *New Ideas in Psychology*, *19*, 237–244.

- Betsch, T., Haberstroh, S., Glöckner, A., Haar, T., & Fiedler, K. (2001). The effects of routine strength on adaptation and information search in recurrent decision making. *Organizational Behavior and Human Decision Processes*, *84*, 23–53.
- Bilalic, M., McLeod, P., & Gobet, F. (2008a). Inflexibility of experts—reality or myth? quantifying the Einstellung effect in chess masters. *Cognitive Psychology*, *56*, 73–102.
- Bilalic, M., McLeod, P., & Gobet, F. (2008b). Why good thoughts block better ones: the mechanism of the pernicious Einstellung (set) effect. *Cognition*, *108*, 652–661.
- Blasi, G. L. (1995). What lawyers know—lawyering expertise, cognitive science, and the function of theory. *Journal of Legal Education*, *45*, 313–397.
- Charness, N., Tuffiash, M., Krampe, R., Reingold, E., & Vasyukova, E. (2005). The role of deliberate practice in chess expertise. *Applied Cognitive Psychology*, *19*, 151–165.
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81.
- Choi, K. S. (1993). Developmental changes in intuitive thinking of Korean children. *Early Child Development and Care*, *85*, 89–95.
- Dickert, S., Herbig, B., Glöckner, A., Gansen, C., & Portack, R. (2012). The more the better?—Effects of training, experience, and information amount in legal judgments. *Applied Cognitive Psychology*, *26*, 223–233.
- Englich, B. (2008). When knowledge matters—differential effects of available knowledge in standard and basic anchoring tasks. *European Journal of Social Psychology*, *38*, 896–904.
- Englich, B., & Soder, K. (2009). Moody experts: how mood and expertise influence judgmental anchoring. *Judgment and Decision Making*, *4*, 41–50.
- Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*, 363.
- Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: in introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise* (pp. 1–38). Cambridge: Cambridge University Press.
- Frensch, P., & Sternberg, R. (1989). Expertise and intelligent thinking: when is it worse to know better. *Advances in the Psychology of Human Intelligence*, *5*, 157–188.
- Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 13124.
- Glöckner, A., & Ebert, I. D. (2011). Legal intuition and expertise. In M. Sinclair (Ed.), *Handbook of Intuition Research* (pp. 157–167). Northampton: Edward Elgar.
- Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2000). Inside the judicial mind. *Cornell Law Review*, *86*, 777–830.
- Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2007). Blinking on the bench: how judges decide cases. *Cornell Law Review*, *93*, 1–44.
- Herbig, B., & Glöckner, A. (2009). Experts and decision making: first steps towards a unifying theory of decision making in novices, intermediates and experts. *MPI Collective Goods Preprint*, No. 2009/2. Available at SSRN: <http://ssrn.com/abstract=1337449>. Accessed 02 Jan 2013.
- Ifenthaler, D., Masduki, I., & Seel, N. M. (2009). The mystery of cognitive structure and how we can detect it: tracking the development of cognitive structures over time. *Instructional Science*, *39*, 1–21.
- Keller, M. C., Fredrickson, B. L., Ybarra, O., Cote, S., Johnson, K., Mikels, J., et al. (2005). A warm heart and a clear head: the contingent effects of weather on mood and cognition. *Psychological Science*, *16*, 724–731.
- Marchant, G., & Robinson, J. (1999). Is knowing the tax code all it takes to be a tax expert? on the development of legal expertise. In R. J. Sternberg & J. S. Horvath (Eds.), *Tacit knowledge in professional practice: researcher and practitioner perspectives* (pp. 3–20). Mahwah: Lawrence Erlbaum.
- Marchant, G., Robinson, J., Anderson, U., & Schadewald, M. (1991). Analogical transfer and expertise in legal reasoning. *Organizational Behavior and Human Decision Processes*, *48*, 272–290.
- Marchant, G., Robinson, J., Anderson, U., & Schadewald, M. (1993). The use of analogy in legal argument: problem similarity, precedent, and expertise. *Organizational Behavior and Human Decision Processes*, *55*, 95–119.
- Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale: Erlbaum.
- Nievelstein, F., van Gog, T., Boshuizen, H. P. A., & Prins, F. J. (2010). Effects of conceptual knowledge and availability of information sources on law students' legal reasoning. *Instructional Science*, *38*, 23–35.
- Patel, V. L., Arocha, J. F., & Kaufman, D. R. (1994). Diagnostic reasoning and medical expertise. *The psychology of learning and motivation: advances in research and theory*, *31*, 187–252.

- Pennington, N., & Hastie, R. (1993). A theory of explanation-based decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsombok (Eds.), *Decision making in action: models and methods* (pp. 188–201). Westport: Ablex.
- Robbenolt, J., MacCoun, R., & Darley, J. (2010). Multiple constraint satisfaction in judging. In D. Klein & G. Mitchell (Eds.), *The psychology of judicial decision making* (pp. 27–40). Oxford: Oxford University Press.
- Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation: cognition in education* (Vol. 55, pp. 1–36). San Diego: Elsevier.
- Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin*, 13, 19–23.
- Schmidt, H., & Boshuizen, H. (1993). On the origin of intermediate effects in clinical case recall. *Memory & Cognition*, 21, 338–351.
- Schweizer, M. (2005). *Kognitive Täuschungen vor Gericht—Eine empirische Studie*. Schweiz: Dissertationsschrift Universität Zürich.
- Simon, H. A., & Chase, W. G. (1973). Skill in chess. *American Scientist*, 61, 393–403.
- Spellman, B. A. (2010). Judges, expertise, and analogy. In D. Klein & G. Mitchell (Eds.), *The psychology of judicial decision making* (pp. 149–163). Oxford: Oxford University Press.
- Sturman, M. C. (2003). Searching for the inverted U-shaped relationship between time and performance: meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management*, 29, 609–640.
- Witteman, C. L. M., & van den Bercken, J. H. L. (2007). Intermediate effects in psychodiagnostic classification. *European Journal of Psychological Assessment*, 23, 56–61.