University of Zurich[UZH]

Department of Business Administration

UZH Business Working Paper Series

Working Paper No. 392

Outcome Bias in Self-evaluations: Quasi-experimental Field Evidence of Swiss Driving License Exams

Pascal Flurin Meier, Raphael Flepp, Philippe Meier and Egon Franck

March 2022

University of Zurich, Plattenstrasse 14, CH-8032 Zurich,
http://www.business.uzh.ch/forschung/wps.html

UZH Business Working Paper Series

Contact Details

Pascal Flurin Meier

pascal-flurin.meier@business.uzh.ch

Raphael Flepp

raphael.flepp@business.uzh.ch

Philippe Meier

philippe.meier@oec.uzh.ch

Egon Franck

egon.franck@business.uzh.ch

University of Zurich

Department of Business Administration

Plattenstrasse 14, CH-8032 Zurich, Switzerland

Outcome Bias in Self-evaluations: Quasi-experimental Field Evidence of Swiss Driving License Exams

Pascal Flurin Meier, Raphael Flepp, Philippe Meier and Egon Franck*

March 1, 2022

Abstract: Employing a quasi-experimental field setting, we examine whether people are outcome biased when self-evaluating their past decisions. Using data from Swiss driving license exams, we find that candidates who narrowly passed the theoretical driving exam are significantly less likely to pass the subsequent practical driving exam – which is taken several months after the theoretical exam – relative to those who failed narrowly. The candidates who passed the theoretical exam in their first attempt received more objections in momentary, on-the-spot kinds of decisions, consistent with the idea that worse preparation is the underlying behavioral difference.

Keywords: Outcome bias; Self-evaluation; Behavioral economics; Judgment; Regression discontinuity design

JEL Classification: D81; D83; D91

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

---

*Corresponding author: Pascal Flurin Meier, University of Zurich, Plattenstrasse 14, 8032 Zurich, Switzerland, pascal-flurin.meier@business.uzh.ch. Coauthors: Raphael Flepp, University of Zurich, Plattenstrasse 14, 8032 Zurich, Switzerland; Philippe Meier, University of Zurich, Plattenstrasse 14, 8032 Zürich; and Egon Franck, University of Zurich, Plattenstrasse 14, 8032 Zurich, Switzerland

1 INTRODUCTION

Do people weight the outcome of a decision more than its informational content when making self-evaluations? For example, consider the decision to drive a car under the influence of alcohol. The driver might evaluate this decision more favorably after an accident-free ride than after a crash under otherwise similar circumstances. Consequently, the driver's future decisions regarding intoxicated driving might be distorted by the outcome of this ride. Such behavior is referred to as outcome bias (Baron & Hershey, 1988).

Prior empirical evidence on outcome bias primarily stems from third-party evaluation settings, where an evaluator assesses the decision of an agent (e.g., Brownback & Kuhn, 2019; Gauriot & Page, 2019; Rubin & Sheremeta, 2016). In contrast, evidence on outcome bias in self-evaluation is scarce and almost exclusively based on laboratory experiments (Jones et al., 1997; Ratner & Herbst, 2005; Tinsley et al., 2012). One notable exception is the field study by Lefgren et al. (2015), who find that professional basketball coaches exhibit outcome bias when revising their starting lineup decisions. However, it remains unclear whether their results are driven by outcome bias in the self-evaluations of coaches or by outcome bias in third-party evaluators (e.g., general managers, fans or media) and unbiased coaches responding accordingly to retain their legitimacy. Overall, even though self-evaluations make up for the lion's share of individuals' real-life decision-making, a thorough understanding of whether outcome bias in self-evaluations distorts subsequent decisions in the field is still lacking.

We fill this research gap by investigating the behavior of young adults in the context of Swiss car driving license exams. This setting offers several unique advantages to study outcome bias in self-evaluation. First, the Swiss car driving licensing system comprises one computer-based theoretical exam and one practical exam on the streets separated by several months, which allows us to test whether outcome bias in self-evaluation after the first exam impacts the results of the second exam. Second, the candidates represent the entire population

of young adults in Switzerland because almost everyone undertakes the driving license process, enabling us to investigate outcome bias in a broad population.

Finally, the passing threshold of the theoretical car driving exam (hereafter "theoretical exam") allows us to compare two groups of candidates who performed very similarly, but one group barely passed the exam while the other group barely failed. Conditional on the number of errors in the theoretical exam around the passing threshold, the outcome of passing or failing is uninformative about the candidate's exam performance Thus, this setting offers a quasi-experimental setting in which the only difference between candidates around the passing threshold is the exam outcome. If candidates exhibit outcome bias in self-evaluation, we expect to observe a difference in the probabilities of passing the practical car driving exam (hereafter "practical exam") between the two groups.

Using data on over 40,000 candidates, we employ a regression discontinuity design (RDD) to test whether self-evaluation and thus the probability of passing the practical exam differs for candidates barely failing and those barely passing the theoretical exam. Our results suggest that candidates who barely failed the theoretical exam in the first attempt are approximately 6 percentage points more likely to pass the practical exam than candidates who barely passed the theoretical exam in the first attempt. This finding is robust to various validation and falsification checks. Thus, candidates who barely failed the theoretical exam seem to self-evaluate their performance more negatively than candidates who barely passed, leading the former to increase their preparation effort for the practical exam relative to the latter.

Despite its advantages, our setting poses a potential concern about the interpretation of our results. Because candidates who barely failed the first theoretical exam need to retake it before being allowed to take the practical exam, they may simply accumulate more theoretical knowledge, which helps them in the practical exam. To distinguish between these alternative

explanations, we make use of our rich data set and draw from the literature on procedural knowledge (Knowlton et al., 2017; Sanchez & Reber, 2013) and skill acquisition (Lewin, 1982). While retaking the theoretical exam might increase declarative knowledge, procedural knowledge of how to drive a car can only be gained through experience and practice. Thus, if outcome bias leads to a different response in the decision regarding how much to prepare for the practical exam, we expect candidates who barely failed the theoretical exam in the first attempt to perform better in evaluation categories that demand more procedural knowledge in the practical exam.

Indeed, we find that candidates who barely failed the theoretical exam have significantly fewer objections in the evaluation categories "traffic visual" and "traffic tactics", which relate to more momentary, on-the-spot kinds of decisions requiring procedural knowledge. However, we fail to find significant differences in evaluation categories relating more to declarative knowledge such as "car control" or "maneuvers".

We add to the literature in several ways. First, our study provides novel field evidence on outcome bias in self-evaluation. In contrast to the study of Lefgren et al. (2015), the source of outcome bias in our setting likely stems from the candidates themselves because third parties such as driving instructors or parents have only limited knowledge about the candidates' theoretical exam performance and preparation level. In addition, we are able to show that outcome bias leads to consequences while Lefgren et al. (2015) only show empirically that it leads to a change in the coaches' strategy. Second, we are the first to document outcome bias in the general population using data that almost perfectly mirror the young Swiss population. This approach thus extends the previous field evidence using sports data involving a subpopulation of individuals acting in a high-stakes environment (e.g., Gauriot & Page, 2019; Kausel et al., 2019; Lefgren et al., 2015). Finally, due to its quasi-experimental features, our

setting allows a clean attribution to a causal effect of the outcome information on subsequent performance.

Our results have important practical implications. First, outcome bias in self-evaluation is present in the (young) population as a whole and is thus not limited to certain subpopulations. Second, there exist many similar institutional settings with multiple separated exams, such as most educational environments or procedures to acquire professional licenses. Thus, outcome bias in self-evaluation might substantially impact the career paths of individuals. Finally, to mitigate the adverse consequences of outcome bias, one plausible solution might be to sensitize candidates that "barely passing the exam" could have easily resulted in "barely failing the exam".

The remainder of this paper is structured as follows. In Section 2, we discuss the related literature. In Section 3, we describe the research design. The results are presented in Section 4, while in Section 5, we address alternative explanations. Section 6 concludes the paper with a discussion.

2 RELATED LITERATURE

First labelled by Baron and Hershey (1988), outcome bias refers to the phenomenon in which people overweight the importance of outcome information in decision-making. Outcomes should be considered in an evaluation only if they provide additional information on the quality of the decision. For instance, if a decision-maker has more information than a judge, then outcomes may be a "valid, albeit imperfect indicator of decision quality" (Hershey & Baron, 1992, p. 90). However, whenever a judge has the same information as a decision-maker ex ante, outcomes are uninformative and should be ignored (Hershey & Baron, 1992). Thus, people are outcome biased when a behavior is considered more justifiable in light of a favorable outcome (Lefgren et al., 2015).

In laboratory experiments, Baron and Hershey (1988) examine subjects who evaluate the appropriateness of third-party decisions, although they also discuss outcome bias in the context of self-evaluation.[1] For instance, their experiments include assessing the appropriateness of a cure (e.g., surgical operation) that had fixed ex ante probabilities (e.g., 8%) of causing severe consequences (e.g., death). The authors consistently show that subjects rated decisions significantly better and the decision-maker more competent when the outcome was favorable than when it was unfavorable.

Most of the subsequent work in laboratory experiments focuses on individuals assessing third-party behavior, such as military combat (Lipshitz, 1989), legal (Alicke et al., 1994) or financial decisions (König-Kersting et al., 2021), ethical judgments (Gino et al., 2010), salesperson performance (Marshall & Mowen, 1993), or audit quality (Peecher & Piercey, 2008), or stems from other principal agent settings and games (e.g., Brownback & Kuhn, 2019; Cushman et al., 2009; Gurdal et al., 2013). Outcome bias is even present in third-party evaluations when the evaluator has complete information on the behavior of the decision-making process, for instance, when knowing the investment strategy of the agent (König-Kersting et al., 2021) or when observing effort (Brownback & Kuhn, 2019).

Field evidence on outcome bias in third-party evaluations mainly stems from sports settings. A notable exception is Emerson et al. (2010), who find a positive outcome bias in peer reviews of evidence-based medicine. Both Gauriot and Page (2019) and Kausel et al. (2019) examine outcome bias utilizing data from professional soccer. Gauriot and Page (2019) exploit the quasi-arbitrary outcome of shots that hit the goal post. It is as good as random whether a long-distance shot hitting the post ultimately lands inside or outside the goal. Hence, the outcome of a long-distance shot on the post does not entail information about the quality of

[1] Baron and Hershey (1988, p. 578) note that people judging their behavior as a function of its outcome "may hold themselves responsible for both good and bad luck, becoming smug in their success or self-reproachful in their failure".

this shot. However, the authors find evidence that lucky post-in shots result in more playing time for the shooter in the upcoming match and in higher ratings by both external journalists and fans than unlucky post-out shots. In a similar spirit, Kausel et al. (2019) examine narrow winnings from penalty shootouts in professional football. The authors find that players on the winning team receive higher ratings from journalists, suggesting that these experts are overly influenced by the outcome of the game. This is true even for players who did not participate in the penalty shootout and therefore could not actively contribute to the outcome of the game after regular and extra time.

In contrast to evidence from third-party evaluations, evidence on outcome bias in self-assessments of the appropriateness of decisions is scarce. However, these kinds of decisions are highly relevant since they are more common in everyday life and have direct personal consequences. In addition, Lefgren et al. (2015) argue that it is more significant to find outcome bias in self-evaluation given that it is easier to be dismissive of others' decisions than one's own. Unlike external judges who act rationally when relying on outcome information in the absence of other information (Baron & Hershey, 1988), individuals who self-evaluate dispose of "process knowledge". They have perfect knowledge about the information available to the decision-maker at the time and about the specific process underlying the decision (Jones et al., 1997). Given these conditions, outcome information is not indicative of decision quality; hence, an effect of outcome information on the judgment of one's own decisions is evidence of bias (Jones et al., 1997).

In the context of self-evaluations, Jones et al. (1997) show that outcome information not only influences participants' evaluations of their decision-making process but also affects their memories in a way to make them consistent with subsequent outcomes. The authors suggest that individuals who experience a good outcome recall their decision-making process as more thoughtful. Ratner and Herbst (2005) find that a negative emotional response resulting from an

unfavorable outcome of a good decision may lead individuals to switch to inferior alternatives. Bachmann (2018) shows in an experiment that advisors can eliminate outcome bias in the context of investment decisions, particularly after bad outcomes. However, advisors are unable to prevent affective reactions after bad outcomes and instead may even reinforce them. Examining decisions in the context of a hazardous situation, Tinsley et al. (2012) find that outcome bias also affects future risk-taking. In addition to conducting laboratory experiments, the authors administer field surveys to residents who assess their actual evacuation decisions following hazard situations. Individuals with resilient near-miss experiences of disasters underestimate the danger of future hazardous situations and increase their risky decision-making (e.g., not engaging in mitigation activities for the potential hazard).

The only field study is Lefgren et al. (2015), who propose that professional coaches exhibit outcome bias when revising their strategy. Utilizing data from top-tier basketball, the authors employ an RDD to study how coaches react to narrow outcomes that are uninformative about team effectiveness or future success and therefore should not impact the coaches' strategy. However, Lefgren et al. (2015) find that coaches are more likely to revise their starting lineup after a narrow defeat than after a narrow win, suggesting that outcome bias is present in the high-stakes decision-making of professionals. The authors acknowledge that they cannot rule out the possibility that the source of outcome bias stems from another stakeholder group, i.e., from a third party, since coaches have to appease a broad set of constituencies (e.g., general manager and fans). As such, the actions of the coaches might be perfectly rational. Alternatively, coaches' actions might also stem from their tendency to favor action over inaction ("bias to action"): Anecdotal evidence suggests that coaches not taking action after losing a game is viewed very unfavorably by fans and other stakeholders. Thus, the clean identification and attribution of outcome bias to a specific stakeholder remains a challenge, particularly in sports settings.

Ultimately, outcome bias can have severe consequences for both evaluated agents and evaluators. Tinsley et al. (2012) suggest that the interpretation of an outcome may lead to different subsequent behaviors in risky decision-making, causing major monetary and emotional costs when an unfavorable outcome occurs. In the context of sports, Lefgren et al. (2015) propose that excessive switching of strategy by coaches matters: It may lead to worse outcomes in future games and impose direct costs, such as putting emotional strain on the actors involved. Flepp and Franck (2021) examine coach dismissal decisions and find that decisions based on factors beyond coaches' control have serious consequences. The authors show that only dismissals following actual poor performance on the pitch improve subsequent team performance, while dismissals after seemingly poor performance (e.g., due to bad luck) do not. Similarly, in the context of CEO turnover decisions, Flepp (2021) shows that boards consider uninformative outcomes outside of the control of the CEO in their decision-making, leading to inefficient and costly dismissal decisions.

In summary, there is a broad strand of literature addressing outcome bias in different contexts. However, evidence on self-evaluation is sparse and predominantly stems from laboratory experiments. In the field, outcome bias and self-evaluations have been examined only in the high-stakes settings of professional team sports (see Lefgren et al., 2015). However, given the natural environment of team sports and its broad set of constituencies, a clean attribution of outcome bias to a specific stakeholder group and thus the generalization to self-judgments is challenging. In the following section, we carefully document how we address outcome bias in the context of self-evaluations of the general young population when applying for the Swiss driving license.

3 RESEARCH DESIGN

3.1 INSTITUTIONAL SETTING & DATA

We utilize a novel setting of young adults applying for a driving license in Switzerland. According to the Federal Statistical Office of Switzerland (2022), 82% of the Swiss population (which was 8.3 million at that time) was allowed to drive a car in 2015. Since a great majority of the Swiss population undertakes the Swiss driving license process at one point, our data closely mirror the general population in their 20s. This is a distinct advantage of our data.

Adults are allowed to drive motor vehicles after they have successfully passed a theoretical exam and a practical exam. The theoretical exam consists of a computer-based multiple-choice exam with 50 questions.[2] A candidate successfully passes the exam if she or he makes fewer than 16 errors. If the candidate fails the exam, she or he can retake it after rescheduling an appointment and paying a registration fee. The agency issues a learner's license valid for 24 months if the exam is successfully passed. This license allows driving a car with an eligible adult as a co-driver and attending driving lessons with a driving instructor. If the student driver feels ready, she or he can register for the practical exam after attending an additional sensitization course consisting of 8 teaching hours.[3] The practical exam consists of 60 minutes driving with an independent official driving expert acting as a co-driver. The expert judges the student driver according to a defined criteria set of evaluation categories. If student driver fails the exam, she or he can retake it in the near future.

Switzerland consists of 26 states, called cantons. The driving license process is consistent across cantons, but each canton is responsible for issuing a driving license for citizens of that canton. We collect data from three different cantons, A, B, and C, which want to remain anonymous. Our sample period starts in 2014 and runs until 2018. We focus on car exams only

[2] It can be scheduled one month prior to the applicant's 18th birthday at the earliest. Applicants who already possess a driving license in a different category (e.g., motorbike) do not have to take the theoretical exam. In our sample, we include only applicants taking the theoretical exam the first time.

[3] Typically, the driving instructor enrolls the student; however, candidates may also self-register.

and on applicants who have no prior driving experience (e.g., a scooter license). We have detailed data at the individual level, which enable us to merge the theoretical exam results with the practical exam results of the candidates. We start with an initial sample of 57,238 candidates for whom we have data on the theoretical exam and the first practical exam. We retain all observations for which we have full data on the first theoretical exam, data on subsequent attempts if the exam was initially failed, and the result of the first practical exam. After we clean the data, the resulting sample includes 44,465 observations containing full data on the theoretical and practical exams.[4]

For the theoretical exam, we observe the number of errors in the first attempt at the theoretical exam (*Errors*). *FirstAttemptFailed* is a dummy variable equaling 1 if the candidate failed the first theoretical exam (i.e., *Errors* equal to or more than 16 errors) and 0 otherwise. We have additional information on gender (*Gender* is a dummy equaling 1 if the candidate is male and 0 otherwise), nationality (*Swiss* is a dummy equaling 1 if the candidate is Swiss and 0 otherwise), and age (*Age* measures age at the time of the theoretical exam).

For our outcome variable, the practical exam, we have rich performance data. On the one hand, we observe whether a candidate passes the exam (*PractExam* is a dummy equaling 1 if the exam is passed and 0 otherwise*)*; on the other hand, we have detailed information about why an examinant failed it. The examiner assesses the candidate based on several factors. If the examiner objects to an action by the candidate, she or he categorizes the violation according to a predefined catalogue consisting of more than 40 possible objections (see Appendix A1 Table A1). If the candidate commits one or more violations, the candidate fails the exam. Candidates who pass the practical exam have zero objections, while candidates who fail the exam have at least 1 objection. We calculate the number of individual objections (*#Objections*) as an alternative outcome variable.

[4] To ensure data validity, we delete observations for which we have strong evidence of flaws (e.g., exams labelled passed with 1 or more objections or duplicates).

In Table 1, we present descriptive statistics of our employed variables. Eighty-six percent of candidates pass the theoretical exam in their first attempt, i.e., have errors below 16. Sixty-four percent of candidates pass the practical exam in their first attempt. The average errors in the first theoretical exam are approximately 8.5. The median candidate is approximately 18.8 years old.

Table 1: Descriptive Statistics

	Definition	N	Mean	Q1	Median	Q3	SD
Dependent variables							
PractExam	1 if the candidate passed the first practical exam	44,465	0.642	0.000	1.000	1.000	0.479
#Objections	Number of objections	44,465	3.316	0.000	0.000	7.000	4.868
Independent variables							
FirstAttemptFailed	1 if candidate failed the theoretical exam in the first attempt	44,465	0.140	0.000	0.000	0.000	0.347
Errors	Number of errors in the first theoretical exam	44,465	8.448	3.000	6.000	11.000	8.065
Age	Age of candidates at theoretical exam	44,465	21.317	18.142	18.764	21.490	5.786
Gender	1 if candidate is male	44,465	0.487	0.000	0.000	1.000	0.500
Swiss	1 if candidate is Swiss	44,465	0.787	1.000	1.000	1.000	0.409

This table reports descriptive statistics for the variables employed. N denotes the number of observations. Mean denotes the mean of the corresponding distribution. The table also displays the median, the 25% quantile (Q1), the 75% quantile (Q3) and the standard deviation (SD) of the distribution.

3.2 EMPIRICAL STRATEGY

We employ an RDD to estimate the causal effect of experiencing different outcomes on future performance. Our quasi-experiment exploits an arbitrary cutoff in the running variable, assigning individuals to treatment or control conditions. The approach relies on the assumption that the characteristics of candidates related to our variables vary smoothly through the threshold. In this case, any discontinuity in future performance may be causally attributed to the treatment, i.e., experiencing a narrow negative outcome.

We test for outcome bias by examining whether candidates who barely failed or barely passed the theoretical exam exhibit different probabilities of passing the practical exam,

conditional on their performance in the first theoretical exam. While a candidate exhibits a positive outcome in the case of narrowly passing the theoretical exam (e.g., with 15 errors), a candidate exhibits a negative outcome in the case of narrow failure (e.g., with 16 errors). Thus, both candidates receive an informative signal about their performance, i.e., the number of errors, suggesting barely sufficient preparation for the exam, and they receive a conditionally uninformative signal of performance, i.e., the narrow outcome of the exam. Crucially, the two types of candidates exhibited almost similar performance, suggesting that their ex ante strategic choice was very similar (e.g., they invested a similar amount of time). If candidates around the threshold have different passing probabilities, it will indicate outcome bias in self-evaluations.

To empirically examine candidates' future performance after experiencing different outcomes, we mainly rely on three sets of results. We first graphically investigate the presence of discontinuity around the threshold. To this end, we plot local sample means of our dependent variable in non-overlapping bins over the number of errors in the theoretical exam. Second, we quantitatively perform our RDD. We rely on local linear regression, i.e., a nonparametric approach, which is frequently used for RD empirical analyses due to its good compromise between flexibility and simplicity (Cattaneo et al., 2019). We test the robustness to our baseline choice and estimate the regression using the whole sample, i.e., the parametric approach. Third, we perform a similar estimation approach with more detailed performance information to more precisely identify the mechanism at stake.

Similar to Klein Teeselink et al. (2021)[5], we use the local linear method proposed by Calonico et al. (2014a) to strike the appropriate balance between bias and precision associated with the use of a local or global RDD approach. Calonico et al. (2014a) implement a nonparametric local polynomial estimation method with optimal bandwidth selection and

[5] Klein Teeselink et al. (2021) employ the method to revisit the phenomena in which losing leads to winning (see also Berger & Pope, 2011).

robust confidence intervals.[6] For an in-depth discussion of the methodology, refer to Calonico et al. (2014b) and Calonico et al. (2019). For the empirical implementation, we utilize the Stata command *rdrobust* developed by the authors (Calonico et al., 2014a; Calonico et al., 2017). Our RD baseline models correspond to the following local regression model where observations are weighted by a kernel function[7] and the bandwidth is data-driven and MSE-optimal:

$$PractExam_i = \alpha + \gamma_0 f(Errors_I - \bar{e}) + \beta\, Treatment_i + \gamma_1 f(Errors_i - \bar{e})\, Treatment_i + \delta\, X_i + \mu_i \quad \textbf{(1)}$$

Treatment is a dummy variable equal to 1 for treated candidates, i.e., those above the threshold, and 0 otherwise (i.e., *FirstAttemptFailed*). Thus, we are interested in the coefficient of β, which equals the treatment effect of failing the theoretical exam on future performance. $PractExam_i$ is a dummy variable indicating whether the candidate successfully passed the subsequent practical exam. f denotes a suitable polynomial function of $Errors_i$, which indicates the number of errors in the theoretical exam, centered at the passing threshold ($\bar{e}$).[8] We allow for possibly different slopes on either side of the threshold; thus, the coefficients on the polynomial terms are indexed by 0 and 1. X_i denotes a vector of covariates. We estimate our baseline specification both with and without covariates. We control for predetermined

[6] The optimal bandwidth selection is based on the data-driven basis of a non-parametric approximation that is a result of a trade-off between the lower variance (associated with a larger bandwidth) and the higher bias (associated with poorer parametric polynomial approximation when using a larger bandwidth) (Calonico et al., 2014a). The authors correct for misspecifications in confidence intervals as a consequence of large bandwidths. They provide a new theory-based and more robust confidence interval estimator for average treatment effects at the cutoff using a bias-corrected RD estimator together with a novel standard error estimator.

[7] We employ a triangular kernel in the baseline model, which gives a positive but declining (linearly and symmetrically) weight to observations within the bandwidth and zero otherwise. When used in conjunction with a bandwidth that optimizes the mean squared error (MSE), it prompts a point estimator with ideal properties (Cattaneo et al., 2019).

[8] For the sake of symmetry in the bandwidth around the threshold, we utilize the threshold of 15.5 errors. We also check the results using the threshold of 16 errors, which leads to similar results.

characteristics such as gender (*Gender*), age (*Age*), and Swiss citizenship (*Swiss*) and include canton fixed effects.[9] Conceptually, covariates are not necessary in RDD but mainly serve to increase precision (Calonico et al., 2019).

To evaluate our results, we check the RD estimates for different bandwidths, polynomial orders and kernel functions in robustness tests. In addition, we estimate equation (1) with an alternative specification of the dependent variable (see Section 4.2). Furthermore, we check for the robustness of the results by employing the methods proposed in Cattaneo et al. (2015) and Cattaneo et al. (2017) and implemented in Cattaneo et al. (2016), which is based on the ideal of local random assignment.

We also estimate equation (1) using logistic regression for the entire sample of candidates, corresponding to the parametric approach (see Section 4.1). Conceptually, the parametric approach equals the nonparametric approach using a uniform kernel with the maximum bandwidth (Cattaneo et al., 2019).

4 RESULTS

4.1 MAIN RESULTS

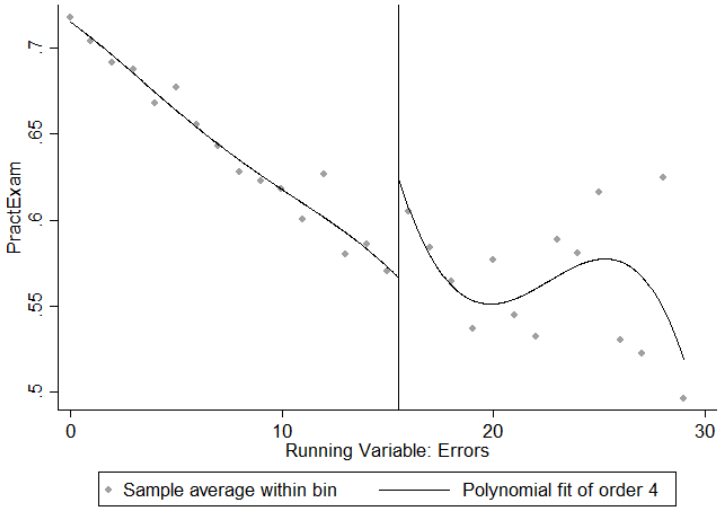
In this section, we present the results of our empirical modeling. First, we illustrate the RDD graphically in Figure 1, which shows a scatterplot with local polynomial fit around the threshold. The clearly visible discontinuity already hints at a positive treatment effect of experiencing a narrow failure: Candidates who experienced a narrow failure on the theoretical exam seem to be more likely to pass the subsequent practical exam than candidates who narrowly passed the theoretical exam.

We next estimate the magnitude of experiencing a narrow failure relative to a narrow pass using both a nonparametric and a parametric approach. The baseline estimates from the

[9] We also include additional examiner-fixed effects, and the results remain virtually the same.

nonparametric approach are reported in Table 2. In addition to the treatment coefficient and the standard error in parentheses, we report the number of observations to the left (#L) and to the right (#R) of the cutoff and the estimated bandwidth for each model in the nonparametric approach (see Table 2 Panel A). We estimate the model using both a conventional RD estimate with a conventional variance estimator and bias-corrected RD estimates with a robust variance estimator, as suggested by Calonico et al. (2014b) and implemented in Calonico et al. (2017).

Figure 1: Graphical Evidence



Notes: The figures shows the regression discontinuity plot for candidates with errors within a bandwidth of 0 to 30 errors. The curves on both sides of the cutoff are fourth-order polynomials.

We find that candidates who narrowly failed the theoretical exam but pass it in a further attempt are more likely to pass the practical driving exam in their first attempt. We interpret this as evidence consistent with outcome bias. Depending on the estimator, candidates who failed the theoretical exam are 5.3 (conventional RD estimator with conventional standard error) or 6.2 (bias-corrected estimates with robust standard errors) percentage points more likely to successfully pass the practical driving exam. Specification II in Panel A of Table 2 includes candidate characteristics as covariates, which should not greatly affect our estimates

if RDD assumptions are valid (Calonico et al., 2019). Indeed, the coefficients remain virtually unchanged.

Table 2: Baseline Results

	Dependent Variable: *PractExam*	
Panel A: Nonparametric Approach	I	II
Conventional Coef.	0.053**	0.056**
	(0.025)	(0.025)
Robust Coef.	0.062**	0.065**
	(0.030)	(0.030)
Bandwidth	4.612	4.416
#L	6477	4859
#R	2911	2481
Covariates	0	5
Panel B: Parametric Approach	I	II
Coef.	0.115***	0.097**
	(0.044)	(0.045)
Observations	44,465	44,465
R-Squared	0.01	0.03
Covariates	0	5

Notes: *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Standard errors are reported in parentheses.
Panel A: Conventional RD estimates with a conventional variance estimator and bias-corrected RD estimates with a robust variance estimator are reported as suggested by Calonico et al. (2014b) and implemented in Calonico et al. (2017). The sample includes observations within the optimal bandwidth selected by one common MSE-optimal bandwidth selector (Calonico et al., 2017). The model is estimated using a triangular kernel and includes a first-degree polynomial, which is allowed to differ on both sides of the cutoff. Model II includes the following covariates: Age, Swiss, Gender and Canton Fixed Effects.
Panel B: The sample includes observations near and far away from the cutoff (parametric approach) weighted by a uniform kernel. The model is estimated using a logistic regression and includes a first-degree polynomial, which is allowed to differ on both sides of the cutoff. Model II includes the following covariates: Age, Swiss, Gender and Canton Fixed Effects.

In the parametric approach, we use observations both close and far from the threshold. In Panel B of Table 2, we present estimates of the parametric approach and use a logistic regression model instead of a linear probability model. We report the baseline results both with and without additional covariates. Again, the results show that the estimated treatment

coefficient is positive and statistically significant, suggesting that candidates who narrowly failed the theoretical car driving exam are more likely to pass the subsequent practical driving exam.[10]

Overall, the results suggest that relative to narrowly passing the theoretical exam, narrowly failing the theoretical exam leads to a considerable increase in the probability of passing the subsequent practical driving exam. These results are consistent with the idea of outcome bias. One plausible explanation follows the idea that the two groups of candidates differ in their amount of preparation. For instance, candidates who passed the theoretical exam narrowly may not prepare as much as candidates who barely failed since they feel more justified about their ex ante decision. We address the potential underlying mechanism in more detail in Section 5.

4.2 VALIDITY TESTS

Local polynomial estimation requires different implementation decisions of the researcher regarding, for example, the polynomial degree, the bandwidth, and the kernel function. In the following, we test the model's sensitivity to our baseline choice. In particular, we use different bandwidths from the optimal bandwidth selection as suggested by Calonico et al. (2017). We start with a bandwidth of 13 to 18 errors and increase it gradually to 8 to 23 errors. We also check the robustness of the baseline results using different kernel functions. We use a uniform kernel, which gives equal weight to all observations within the bandwidth, and the Epanechnikov kernel, which gives a quadratic decaying weight to observations within the bandwidth (Cattaneo et al., 2019). In addition, we test the sensitivity to the polynomial degree by utilizing higher-order polynomials than our default choice of the first-order degree.[11]

[10] The results also remain the same when we include additional examiner fixed effects.
[11] Following the authors, higher-order polynomials tend to produce overfitting and lead to unreliable results near boundary points, which, together with the increasing variability of the treatment effect estimator, lead the local linear estimator to be the preferred point estimator in many applications (Cattaneo et al., 2019).

Again, we report both conventional RD estimates with a conventional variance estimator and bias-corrected RD estimates with a robust variance estimator. The results are displayed in

Table 3.

Overall, our estimates are not sensitive to our varying the bandwidth, at least in terms of effect size, while statistical significance is not present in all specifications, particularly for narrow bandwidths. Additionally, our results are not sensitive to the choice of the kernel function, while they are somewhat sensitive to polynomials of order three, at least in terms of statistical significance.

Given that our main outcome variable is binary, to further substantiate our results, we utilize the number of objections (*#Objections*) as an alternative outcome variable. If a candidate passes the practical exam, the variable equals zero. Thus, under these specifications, we expect a negative effect. The results are displayed in Table 4. The results are robust under both estimators and in line with the baseline results, indicating that candidates who experience a narrow failure on the theoretical exam receive fewer objections overall. The results confirm our baseline results that candidates who failed narrowly at first not only have a higher probability of successfully passing the practical component but also receive fewer objections.

As a last robustness test, we show evidence from a different RD approach proposed in Cattaneo et al. (2015) and Cattaneo et al. (2017) and implemented in Cattaneo et al. (2016), which is based on the ideal of local random assignment. The results are significant and comparable in size of the effect (see Appendix A2 Table A2).

Table 3: Robustness Tests

Dependent Variable: *PractExam*

Panel A: **Bandwidth**	I	II	III	IV	V	VI
Coef.	0.056	0.048	0.056**	0.046**	0.040*	0.035*
Conventional	(0.037)	(0.029)	(0.025)	(0.023)	(0.021)	(0.019)
Coef.	0.066	0.065	0.041	0.066*	0.064**	0.059**
Robust	(0.054)	(0.054)	(0.041)	(0.036)	(0.032)	(0.029)
Kernel	Trian.	Trian.	Trian.	Trian.	Trian.	Trian.
Polynomial Degree	1	1	1	1	1	1
Intervall	**13-18**	**12-19**	**11-20**	**10-21**	**9-22**	**8-23**
#L	3353	4859	6477	8298	10420	12816
#R	1974	2481	2911	3282	3605	3875
Covariates	5	5	5	5	5	5
Panel B: **Kernel function**	I	II	III	IV	V	VI
Coef.	0.053**	0.056**	0.061**	0.062**	0.056**	0.056**
Conventional	(0.025)	(0.025)	(0.024)	(0.024)	(0.025)	(0.025)
Coef.	0.062**	0.065**	0.072**	0.071**	0.064**	0.066**
Robust	(0.030)	(0.030)	(0.028)	(0.029)	(0.030)	(0.030)
Kernel	**Trian.**	**Trian.**	**Uni.**	**Uni.**	**Epa.**	**Epa.**
Polynomial Degree	1	1	1	1	1	1
Bandwith	4.612	4.416	3.753	3.552	4.287	4.179
#L	6477	4859	4859	4859	4859	4859
#R	2911	2481	2481	2481	2481	2481
Covariates	**0**	**5**	**0**	**5**	**0**	**5**
Panel C: **Polynomial degree**	I	II	III	IV	V	VI
Coef.	0.053**	0.057**	0.061	0.056**	0.059**	0.063
Conventional	(0.025)	(0.028)	(0.040)	(0.025)	(0.028)	(0.038)
Coef.	0.062**	0.064**	0.060	0.065**	0.067**	0.060
Robust	(0.030)	(0.032)	(0.047)	(0.030)	(0.032)	(0.044)
Kernel	Trian.	Trian	Trian	Trian	Trian	Trian
Polynomial Degree	**1**	**2**	**3**	**1**	**2**	**3**
Bandwidth	4.612	7.807	7.694	4.416	7.747	8.375
#L	6477	12816	12816	4859	12816	12816
#R	2911	3875	3875	2481	3875	3875
Covariates	**0**	**0**	**0**	**5**	**5**	**5**

Notes: *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Standard errors are reported in parentheses. Conventional RD estimates with a conventional variance estimator and bias-corrected RD estimates with a robust variance estimator are reported as suggested by Calonico et al. (2014b) and implemented in Calonico et al. (2017). The baseline specification includes observations within the optimal bandwidth selected by one common MSE-optimal bandwidth selector (Calonico et al., 2017). The baseline model is estimated using a triangular kernel and includes a first-degree polynomial, which is allowed to differ on both sides of the cutoff.

Table 4: Alternative Dependent Variable

		Dependent Variable: *#Objections*	
		I	II
Conventional	Coef.	-0.487*	-0.562**
		(0.256)	(0.255)
Robust	Coef.	-0.579*	-0.661**
		(0.308)	(0.306)
	Bandwidth	4.907	4.694
	#L	6477	6477
	#R	2911	2911
	Covariates	0	5

Notes: *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Standard errors are reported in parentheses. Conventional RD estimates with a conventional variance estimator and bias-corrected RD estimates with a robust variance estimator are reported as suggested by Calonico et al. (2014b) and implemented in Calonico et al. (2017). The sample includes observations within the optimal bandwidth selected by one common MSE-optimal bandwidth selector (Calonico et al., 2017). The model is estimated using a triangular kernel and includes a first-degree polynomial, which is allowed to differ on both sides of the cutoff. Model II includes the following covariates: Age, Swiss, Gender and Canton Fixed Effects.

The identifying assumption requires that candidates be randomly allocated above and below the threshold and may not manipulate the treatment condition. Thus, the candidates who narrowly pass the theoretical driving license exam are likely to be similar to candidates who narrowly fail the exam in terms of their individual characteristics and preparation for the exam. This assumption is feasible for several reasons. First, candidates have no prior experience of testing at this level given that it is a completely new task and context. They are unlikely to exactly predict the amount of preparation needed to pass the theoretical exam (narrowly). Second, similar to Proud (2015), we argue that even if they were able to predict the amount of effort needed, their realized errors – which are immediately communicated since the exam is automatically evaluated – are still expected to be random around the threshold due to random shock terms.

One remaining concern with an RDD specification is that the treatment group is systematically different than the control group. As a consequence, one would expect different levels of performance in the subsequent practical exam, depending on individual

characteristics. To mitigate such concerns, we use the same local polynomial regression model (1) for our predetermined covariates. As highlighted by Cattaneo et al. (2019), we analyze each covariate as if it were an outcome, including the choice of an optimal bandwidth and performing local polynomial inference within that bandwidth. If there were evidence for the null hypothesis of no treatment to reject, it would question the validity of the design. Table 5 shows no evidence to reject the null hypothesis of no treatment.

Table 5: Balance checks

Dependent Variable: ***Predetermined Covariates***			
Balance checks	Age	Gender	Swiss
Coef.	0.124	-0.036	0.023
Conventional	(0.249)	(0.026)	(0.017)
Coef.	0.180	-0.046	0.027
Robust	(0.300)	(0.031)	(0.021)
Bandwidth	6.628	4.379	7.973
#L	10420	4859	12816
#R	3605	2481	3875

Notes: *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Standard errors are reported in parentheses. Conventional RD estimates with a conventional variance estimator and bias-corrected RD estimates with a robust variance estimator are reported as suggested by Calonico et al. (2014b) and implemented in Calonico et al. (2017). The sample includes observations within the optimal bandwidth selected by one common MSE-optimal bandwidth selector (Calonico et al., 2017). The model is estimated using a triangular kernel and includes a first-degree polynomial, which is allowed to differ on both sides of the cutoff.

The key identifying assumption relies on the continuity of the regression functions for treatment and control units at the cutoff in the absence of the treatment. Unfortunately, this condition is fundamentally untestable. However, a useful falsification analysis of the RDD, in addition to the balance checks of covariates, is to examine a treatment effect at artificial or placebo thresholds. Thus, we test whether the regression function is continuous at thresholds other than the actual treatment cutoff. Evidence of continuity away from the actual cutoff is not a necessary or a sufficient condition at the cutoff. However, discontinuities away from the actual cutoff raise questions of validity of the RDD (Cattaneo et al., 2019). Thus, we examine

the neighboring artificial cutoffs around 14, 15, 17 and 18 errors to check for discontinuities apart from the actual treatment threshold at 16 errors. In addition, we use the artificial threshold of plus-minus 5 errors, i.e., 11 and 21. The results are presented in Table 6. We see no evidence of discontinuity using the artificial thresholds, which leaves us confident that our identification strategy is valid.

Table 6: Falsification Test

Dependent Variable: *PractExam*							
	I	II	III	IV	V	VI	VII
Coef.	-0.009	-0.006	0.003	0.053**	-0.001	-0.026	-0.019
Conventional	(0.021)	(0.022)	(0.021)	(0.025)	(0.023)	(0.022)	(0.030)
Coef.	-0.017	-0.009	-0.001	0.062**	0.006	-0.031	-0.032
Robust	(0.025)	(0.027)	(0.026)	(0.030)	(0.028)	(0.027)	(0.034)
Threshold	**11**	**14**	**15**	*16*	**17**	**18**	**21**
Bandwidth	3.399	4.451	5.148	4.612	5.977	7.835	6.658
#L	6339	6222	7328	6477	7241	9726	4987
#R	4401	3504	3451	2911	2841	2918	1798
Covariates	0	0	0	*0*	0	0	0

Notes: *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Standard errors are reported in parentheses. Conventional RD estimates with a conventional variance estimator and bias-corrected RD estimates with a robust variance estimator are reported as suggested by Calonico et al. (2014b) and implemented in Calonico et al. (2017). The sample includes observations within the optimal bandwidth selected by one common MSE-optimal bandwidth selector (Calonico et al., 2017). The model is estimated using a triangular kernel and includes a first-degree polynomial, which is allowed to differ on both sides of the cutoff.

5 ALTERNATIVE EXPLANATION

Our baseline results suggest that candidates who narrowly passed the theoretical exam are more likely to pass the subsequent practical exam. This result is consistent with the explanation of outcome bias. However, the institutional setting also provides a potential alternative explanation. Since candidates who fail the first theoretical exam need to retake it before being allowed to take the practical exam, those who barely failed may simply accumulate more theoretical knowledge, which helps them in the practical exam, than those who barely passed the theoretical exam.

To test this alternative explanation empirically[12], we make use of our rich data on individual objections in the practical exam, categorized according to the official classification of the Road Traffic Department of each canton. The objection categories are handling, traffic vision, traffic environment, traffic tactics, traffic dynamics, maneuver and car control. If the accumulation of knowledge is the main mechanism, we expect to observe a different pattern among candidates who barely failed or passed the theoretical exam.

Each category subsumes multiple individual objections; thus, the seven categories of objections consist of different individual objections.[13] Cantons A and B utilize the same classification list, and canton C uses a similar but slightly different list.[14] A full list of the individual objections as well as their categorizations and short descriptions is presented in Appendix A1 Table A1. We consider that objection categories (e.g., maneuver) subsume multiple individual objections (e.g., parking, reversing, or turning).

To theoretically connect the objections to the accumulation of knowledge as an alternative explanation, we draw from the literature on procedural knowledge and learning (Knowlton et al., 2017) and on the acquisition of skills (Lewin, 1982). Learning how to drive a car is an example of the acquisition of a complex perceptual-motor skill (Fitts, 1964, 1965; Fitts & Posner, 1967; Fleishman, 1965; Lewin, 1982).[15] In the associative stage of skill acquisition (i.e., refining and coordination by experience and practice), "procedural

[12] Before taking the practical exam, candidates have to attend an additional basic theoretical sensitization course, consisting of eight obligatory lessons. This might support the fact that candidates are on a comparable level in terms of theoretical knowledge.

[13] For instance, if a candidate fails to properly park backwards, wrongly performs an emergency brake and overlooks a right of way, the examiner marks objections in parking, emergency brake and right of way, respectively. The candidate fails the practical exam as he or she has three objections. Each individual objection is classified in one broader category; while parking and emergency brake refer to a wrongfully performed *maneuver*, right of way refers to *traffic tactics* (see Appendix A1 Table A1).

[14] The categories of these objections are the same for all three cantons, except for an additional objection for cantons A and B, which we refer to as "miscellaneous" (a pool of objections that are not classified). While the majority of the objections overlap, some objections are unique and others are classified differently. For instance, cantons A and B have only one objection for parking, while canton C differentiates between parking sideways, forwards and backwards.

[15] Candidates start by acquiring declarative knowledge ("the cognitive stage"), which is assessed through a theoretical driving exam. Then, they proceed to actually learning how to drive while building up procedural knowledge ("the associative stage").

knowledge" is being acquired, representing procedures gained through experience (e.g., accelerate when entering a highway) or "knowing how" (Knowlton et al., 2017). It contrasts with the declarative memory, or "knowing that" (e.g., the speed limit). The concern comes down to whether it is plausible that by retaking the theoretical exam, the candidate accumulates more declarative knowledge, which facilitates the acquisition of procedural knowledge and ultimately improves performance on the practical exam.[16]

If our results are driven by the accumulation of theoretical knowledge, we expect to observe a stronger effect for objections that are more strongly associated with declarative knowledge than with procedural knowledge. We expect declarative knowledge to be more important in objections classified under *maneuver* or *car handling*. These objections require knowing how the operating equipment works, positioning the vehicle, or performing a time-pressure-free maneuver such as parking backwards, which follows a typical learned manual. Conversely, if the results are indeed driven by outcome bias, the amount of preparation for the practical exam likely differs between the two groups of candidates. This should translate into a distinct pattern for objections in which candidates who barely failed the theoretical exam receive fewer objections in categories that are more associated with experience.

Examining the impact of attention on sensorimotor skills in novices and experts, Beilock et al. (2002) find that the performance of novices is impaired in environments that divert attention away from the primary tasks, whereas more experienced performers are not adversely affected. In contrast, novices even benefit from conditions that prompt their attention to the task properties, while experienced agents are harmed when they must devote explicit attention to skills processes that normally run automatically.[17] As procedural knowledge develops

[16] Knowlton et al. (2017) argue that in many cases, both types of memories make contributions and that some declarative knowledge is needed as a foundation of procedural learning.

[17] The authors also suggest that if some aspects of the skill are not as well developed, then experienced agents may also benefit from conditions that prompt their attention to the task at hand rather than being distracted (Beilock et al., 2002).

through practice, skills that rely on it can be executed without as much deliberate attention and be more readily accessible under distracting circumstances. This may help candidates direct their attention to other stimuli and be better able to handle conditions in create dual-task environments (Beilock et al., 2002), such as those often encountered on the road.

We expect such dual-task environments to be particularly present in the *traffic visuals, traffic dynamics,* and *traffic tactics* categories. These categories require substantial attentional resources since they grasp dynamics on the road, such as interactions with other road users, which rely on hands-on experience on the road. For the empirical implementation, we calculate an indicator variable equaling 1 if the candidate violates at least 1 aspect in an objection category and 0 otherwise. Thus, we calculate the following indicator variables: *Maneuver, Visual, Dynamics, Environment, Handling, Tactics* and *Control.* We fit the same baseline estimation model (1) using the seven indicator variables.[18]

Table 7: Categories of Objections

	Dependent Variable: *Indicator Variable*						
Panel A: Categories	I Maneuver	II Visual	III Dynamics	IV Environ.	V Handling	VI Tactics	VII Control
Coef.	-0.023	-0.074***	-0.028	-0.003	-0.002	-0.062**	-0.034
Conventional	(0.021)	(0.025)	(0.021)	(0.011)	(0.016)	(0.025)	(0.023)
Coef.	-0.030	-0.086***	-0.036	-0.006	-0.006	-0.072**	-0.040
Robust	(0.026)	(0.030)	(0.024)	(0.013)	(0.019)	(0.030)	(0.028)
Bandwidth	4.580	4.142	5.621	5.161	5.172	4.364	4.541
#L	6477	4859	8298	6477	6477	4859	6477
#R	2911	2481	3282	2911	2911	2481	2911
Covariates	5	5	5	5	5	5	5

Notes: *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Standard errors are reported in parentheses. Conventional RD estimates with a conventional variance estimator and bias-corrected RD estimates with a robust variance estimator are reported as suggested by Calonico et al. (2014b) and implemented in Calonico et al. (2017). The sample includes observations within the optimal bandwidth selected by one common MSE-optimal bandwidth selector (Calonico et al., 2017). The model is estimated using a triangular kernel and includes a first-degree polynomial, which is allowed to differ on both sides of the cutoff. All models include the following covariates: Age, Swiss, Gender and Canton Fixed Effects.

[18] Our conclusions remain similar when we consider the number of objections per category.

The results presented in Panel A of Table 7 suggest that candidates who experienced an initial setback are likely to receive significantly less objections in *traffic visuals* and *traffic tactics*. The coefficients are highly significant and largest in terms of size. Overall, these results are consistent with the explanation of outcome bias. The results support the idea that candidates who experienced a narrow failure show a higher preparation level than candidates who experienced a narrow success.

6 DISCUSSION AND CONCLUSION

In this paper, we present novel field evidence suggesting that outcome bias is present when individuals self-evaluate their behavior. We find that candidates who narrowly passed the theoretical car driving exam in their first attempt have a significantly lower probability of passing the practical car driving exam compared to those who barely failed the theoretical exam. Examining more detailed performance data for the practical car driving exam, we find that candidates who narrowly failed the theoretical exam receive fewer objections in their momentary, on-the-spot decisions, suggesting they might be better able to handle dual-task environments due to experience. We interpret these findings as evidence consistent with the notion of outcome bias, suggesting that candidates are likely to differ in their amount of preparation for the practical car driving exam. We cannot precisely say whether candidates who barely passed the exam or candidates who barely failed the exam exhibit this tendency. However, our results indicate that at least one group falls prey to outcome bias. Taken together, our results suggest that outcome bias in self-evaluation is indeed a generalizable phenomenon.

In the context of our institutional setting, outcome bias has consequences. On the one hand, candidates must wait to finally be able to drive a car – which, for young people at this age, is an important life event and a signal. On the other hand, outcome bias leads to considerable monetary costs. Conservatively assuming, the extra costs of an additional attempt are at least USD 141 in the case without additional driving lessons but up to approximately

USD 900 in the case with five additional driving lessons and the provision of the driving instructor's car.[19] The majority of young adults aged 18 to 20 are typically in their last year of vocation training, earning a median gross income of USD 1,360 per month (Federal Statistical Office, 2021). Thus, in contrast to Lefgren et al. (2015), we show empirically that outcome bias in self-evaluations leads to actual consequences.

One limitation of our institutional setting is that we cannot cleanly observe the behavioral adjustments made by both groups. Instead, we observe only their subsequent outcomes. By examining detailed performance data and relying on theory on the acquisition of skills and procedural knowledge, we try to minimize the concern that the accumulation of theoretical knowledge as an alternative explanation underpins the main effect. In a related spirit, we cannot infer to what extent both groups adjust their strategy. We only observe a relative difference between the two groups. Thus, we cannot conclude that candidates who barely passed the theoretical exam do not adjust their strategy; we can only conclude with reasonable certainty that they do less so than candidates who failed narrowly.

Taken together, our results suggest that outcome bias is indeed a rather general phenomenon and is not limited to a subpopulation. It applies to a variety of different contexts in real life and is important given the numerous decisions involving self-evaluations people face each day. Since there are many similar institutional settings with multiple separated exams (e.g., most educational environments and acquisitions of professional licenses), outcome bias in self-evaluation might substantially impact the career paths of individuals. Finally, to mitigate

[19] The underlying assumptions are as follows (CHF 1 = USD 1.09 as of October 26, 2021):

1) USD 140 — Fee for the practical exam (CHF 130)
2) USD 765 — Costs for seven additional driving lessons (excluding lessons with family, friends and fools) Assumptions: 50 days on average between the exams, assuming one lesson per 10 days and provision of the car of the driving instructor, assuming two additional lessons. The price of a lesson depends on the canton but can easily go beyond USD 109 (CHF 100) (per 45 to 50 minutes) in the most expensive cantons.

adverse consequences of outcome bias, one plausible solution might be to sensitize candidates that "barely passing the exam" could easily have resulted in "barely failing the exam".

We believe that there is room for future research on outcome bias in the context of self-evaluation. Since it is crucial to understand how to enhance decision-making other than simply by raising awareness, examining factors or conditions to overcome outcome bias in self-evaluations would be a fruitful avenue for further research. To date, our understanding is still limited. In the context of third-party evaluation, Sezer et al. (2016) find that outcome bias is reduced under separate evaluation relative to joint evaluation, suggesting that joint evaluation makes attending to information about intention more difficult. Gillenkirch and Velthuis (2018) show that outcome bias is stronger when peer comparison is present. However, it remains to be explored which factors amplify or limit this bias, as other conditions and factors might prevail when individuals self-evaluate their decisions. For instance, in the context of self-evaluating investment decisions, Bachmann (2018) shows that advising can eliminate outcome bias by eliminating uncertainty in the quality of decisions, particularly after bad outcomes.

REFERENCES

Alicke, M. D., Davis, T. L., & Pezzo, M. V. (1994). A posteriori adjustment of a priori decision criteria. *Social Cognition*, *12*(4), 281–308. https://doi.org/10.1521/soco.1994.12.4.281

Bachmann, K. (2018). Can advisors eliminate the outcome bias in judgements and outcome-based emotions? *Review of Behavioral Finance*, *10*(4), 336–352. https://doi.org/10.1108/RBF-11-2016-0072

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569. https://doi.org/10.1037/0022-3514.54.4.569

Beilock, S. L., Carr, T. H., MacMahon, C., & Starkes, J. L. (2002). When paying attention becomes counterproductive: impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology: Applied*, *8*(1), 6. https://doi.org/10.1037/1076-898X.8.1.6

Berger, J., & Pope, D. (2011). Can losing lead to winning? *Management Science*, *57*(5), 817–827. https://doi.org/10.1287/mnsc.1110.1328

Brownback, A., & Kuhn, M. A. (2019). Understanding outcome bias. *Games and Economic Behavior*, *117*, 342–360. https://doi.org/10.1016/j.geb.2019.07.003

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R [Rocio] (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, *17*(2), 372–404. https://doi.org/10.1177/1536867X1701700208

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R [Rocio] (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, *101*(3), 442–451. https://doi.org/10.1162/rest_a_00760

Calonico, S., Cattaneo, M. D., & Titiunik, R [Rocio] (2014a). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal*, *14*(4), 909–946. https://doi.org/10.1177/1536867X1701700208

Calonico, S., Cattaneo, M. D., & Titiunik, R [Rocio] (2014b). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326. https://doi.org/10.3982/ECTA11757

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R [Rocio] (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, *3*(1), 1–24. https://doi.org/10.1515/jci-2013-0010

Cattaneo, M. D., Idrobo, N., & Titiunik, R [Rocío] (2019). A practical introduction to regression discontinuity designs: Foundations. Advance online publication. https://doi.org/10.1017/9781108684606

Cattaneo, M. D., Titiunik, R [Rocio], & Vazquez-Bare, G. (2016). Inference in regression discontinuity designs under local randomization. *The Stata Journal*, *16*(2), 331–367. https://doi.org/10.1177/1536867X1601600205

Cattaneo, M. D., Titiunik, R [Rocio], & Vazquez-Bare, G. (2017). Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management*, *36*(3), 643–681. https://doi.org/10.1002/pam.21985

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS One*, *4*(8), e6699. https://doi.org/10.1371/journal.pone.0006699

Emerson, G. B., Warme, W. J., Wolf, F. M., Heckman, J. D., Brand, R. A., & Leopold, S. S. (2010). Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. *Archives of Internal Medicine*, *170*(21), 1934–1939. https://doi.org/10.1001/archinternmed.2010.406

Federal Statistical Office. (2021, July 16). *Bruttoerwerbseinkommen pro Jahr der Erwerbstätigen nach Erwerbsstatus, Berufsgruppen ISCO 08, Beschäftigungsgrad und Geschlecht - Zentralwert (Median) in Franken*. https://www.bfs.admin.ch/bfs/de/home/statistiken/arbeit-erwerb/loehne-erwerbseinkommen-arbeitskosten.assetdetail.17604449.html

Federal Statistical Office. (2022, February 21). *Besitz von Fahrzeugen, Führerausweisen und ÖV-Abos*. https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/personenverkehr/verkehrsverhalten/besitz-fahrzeuge-fahrausweise.html

Fitts, P. M. (1964). Perceptual-motor skill learning. In *Categories of human learning* (pp. 243–285). Elsevier.

Fitts, P. M. (1965). Factors in complex skill training. In R. Glaser (Ed.), *Training research and education* (Vol. 1962, pp. 177–197). J. Wiley.

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole.

Fleishman, E. A. (1965). The description and prediction of perceptual-motor skill learning. In R. Glaser (Ed.), *Training research and education* (pp. 137–175). J. Wiley.

Flepp, R. (2021). Uninformative performance signals and forced CEO turnover. *Available at SSRN 3904056.* Advance online publication. https://doi.org/10.2139/ssrn.3904056

Flepp, R., & Franck, E. (2021). The performance effects of wise and unwise managerial dismissals. *Economic Inquiry*, *59*(1), 186–198. https://doi.org/10.1111/ecin.12924

Gauriot, R., & Page, L. (2019). Fooled by performance randomness: overrewarding luck. *Review of Economics and Statistics*, *101*(4), 658–666. https://doi.org/10.1162/rest_a_00783

Gillenkirch, R. M., & Velthuis, L. (2018). Subjective evaluations of risk taking decisions: experimental evidence on outcome biases and their consequences. *Available at SSRN 3232703.* Advance online publication. https://doi.org/10.2139/ssrn.3232703

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless+ harmless= blameless: When seemingly irrelevant factors influence judgment of (un) ethical behavior. *Organizational Behavior and Human Decision Processes*, *111*(2), 93–101. https://doi.org/10.1016/j.obhdp.2009.11.001

Gurdal, M. Y., Miller, J. B., & Rustichini, A. (2013). Why blame? *Journal of Political Economy*, *121*(6), 1205–1247. https://doi.org/10.1086/674409

Hershey, J. C., & Baron, J. (1992). Judgment by outcomes: When is it justified? *Organizational Behavior and Human Decision Processes*, *53*(1), 89–93. https://doi.org/10.1016/0749-5978(92)90056-D

Jones, S. K., Yurak, T. J., & Frisch, D. (1997). The effect of outcome information on the evaluation and recall of individuals' own decisions. *Organizational Behavior and Human Decision Processes*, *71*(1), 95–120. https://doi.org/10.1006/obhd.1997.2714

Kausel, E. E., Ventura, S., & Rodríguez, A. (2019). Outcome bias in subjective ratings of performance: Evidence from the (football) field. *Journal of Economic Psychology*, *75*, 102132. https://doi.org/10.1016/j.joep.2018.12.006

Klein Teeselink, B., van den Assem, Martijn J, & van Dolder, D. (2021). Does losing lead to winning? An empirical analysis for four different sports. *Management Science, Forthcoming*. Advance online publication. https://doi.org/10.2139/ssrn.3669174

Knowlton, B., Siegel, A., & Moody, T. (2017). Procedural learning in humans. In https://doi.org/10.1016/B978-0-12-809324-5.21085-7

König-Kersting, C., Pollmann, M., Potters, J., & Trautmann, S. T. (2021). Good decision vs. good results: Outcome bias in the evaluation of financial agents. *Theory and Decision*, *90*(1), 31–61. https://doi.org/10.1007/s11238-020-09773-1

Lefgren, L., Platt, B., & Price, J. (2015). Sticking with what (barely) worked: A test of outcome bias. *Management Science*, *61*(5), 1121–1136. https://doi.org/10.1287/mnsc.2014.1966

Lewin, I. (1982). Driver training: a perceptual-motor skill approach. *Ergonomics*, *25*(10), 917–924. https://doi.org/10.1080/00140138208925051

Lipshitz, R. (1989). "Either a medal or a corporal": The effects of success and failure on the evaluation of decision making and decision makers. *Organizational Behavior and Human Decision Processes*, *44*(3), 380–395. https://doi.org/10.1016/0749-5978(89)90015-0

Marshall, G. W., & Mowen, J. C. (1993). An experimental investigation of the outcome bias in salesperson performance evaluations. *Journal of Personal Selling & Sales Management*, *13*(3), 31–47.

Peecher, M. E., & Piercey, M. D. (2008). Judging audit quality in light of adverse outcomes: Evidence of outcome bias and reverse outcome bias. *Contemporary Accounting Research*, *25*(1), 243–274. https://doi.org/10.1506/car.25.1.10

Proud, S. (2015). Resits in higher education: merely a bar to jump over, or do they give a pedagogical 'leg up'? *Assessment & Evaluation in Higher Education*, *40*(5), 681–697. https://doi.org/10.1080/02602938.2014.947241

Ratner, R. K., & Herbst, K. C. (2005). When good decisions have bad outcomes: The impact of affect on switching behavior. *Organizational Behavior and Human Decision Processes*, *96*(1), 23–37. https://doi.org/10.1016/j.obhdp.2004.09.003

Rubin, J., & Sheremeta, R. (2016). Principal–agent settings with random shocks. *Management Science*, *62*(4), 985–999. https://doi.org/10.1287/mnsc.2015.2177

Sanchez, D. J., & Reber, P. J. (2013). Explicit pre-training instruction does not improve implicit perceptual-motor sequence learning. *Cognition*, *126*(3), 341–351. https://doi.org/10.1016/j.cognition.2012.11.006

Sezer, O., Zhang, T., Gino, F., & Bazerman, M. H. (2016). Overcoming the outcome bias: Making intentions matter. *Organizational Behavior and Human Decision Processes*, *137*, 13–26. https://doi.org/10.1016/j.obhdp.2016.07.001

Tinsley, C. H., Dillon, R. L., & Cronin, M. A. (2012). How near-miss events amplify or attenuate risky decision making. *Management Science*, *58*(9), 1596–1613. https://doi.org/10.1287/mnsc.1120.1517

Appendix A1

Table A1: Objections Overview

Objection	Description	Canton	Category
Launching	*Coupling, all-round check, brake test*	A / B / C	Handling
Operating equipment	*Armatures, operating equipment, lights, blinker, tachograph, etc.*	A / B / C	Handling
Positioning	*Wheel, seat, headrest, mirrors, positioning*	A / B / C	Handling
Basic operations	*Accelerate, different brakes, clutch, shifting, gear selection*	A / B / C	Handling
Familiarity	*Ride comfort, pillion rides*	A / B / C	Handling
Viewing systematics	*Order, double and side view, follow-up, appropriateness*	A / B / C	Visual
Foresight	*Recognition of danger spots, behavior, defensive driving*	A / B / C	Visual
Sensoring	*Road users, weather, road*	A / B / C	Visual
Viewing technique	*Curves, bottlenecks, bends, branches*	A / B / C	Visual
Orientation	*Front, back, side, mirrors*	A / B / C	Visual
Partnering	*3A, communication, consideration*	A / B / C	Environ.
Conditions	*Road, visibility and weather conditions*	A / B / C	Environ.
Speed	*Keep up, differentiate, adjust, exceed*	A / B / C	Dynamics
Movement	*Communication, lane, spurt*	A / B / C	Dynamics
Road use	*Road partitioning, lane keeping, driving on the right*	A / B / C	Dynamics
Curves	*Left, right, cutting inside turns, sequence, line selection, speed*	A / B / C	Dynamics
Changing lanes	*Use of gaps, turning left/right, sequence, observation, obstruction, control*	A / B / C	Dynamics
Passing	*Crossing, passing, overtaking, distance, obstruction, time, speed, allowed*	A / B / C	Dynamics
Keeping up	*Column driving, side by side, behind each other, distance*	A / B / C	Dynamics
Partnering maneuver	*Passing by, use of gaps, keeping up, highway*	C	Dyanamics
Driving physics	*Stopping, driving stability*	C	Dynamics
Roundabouts	*Navigation, dynamics, use of gaps*	A / B / C	Tactics
Signalling	*Presence, timing*	A / B / C	Tactics
Entering roads	*Sequence, with/without lane, one-way, opposite lane, tram tracks*	A / B / C	Tactics
Braking readiness	*Presence, timing, appropirateness, wrong*	A / B / C	Tactics
Precedence	*Safety, disregard, waiver, signalization, precedence from right, unsecure, handling*	A / B / C	Tactics
Traffic signals	*Unsecure, disregard, following, respecting (traffic signals, signposts, light signals, markings, police, traffic regulation)*	A / B / C	Tactics
Pedestrians	*Precedence, behavior, disregard*	A / B / C	Tactics
Public transport	*Bus, tram, railway crossing*	A / B	Tactics
Highway	*Entry and exit, speed, distance, overtaking, driving on the right*	A / B / C	Tactics
Obstruction	*Obstruction*	A / B / C	Control
Hazards	*Abstract, likely, concrete*	A / B / C	Control
Intervention	*Verbal, steering, brake, gas*	A / B / C	Control
Abortion	*Abortion of exam*	A / B / C	Control
Stopping / starting	*Gradient, slope, observation*	A / B / C	Maneuver
Parking	*Left, right, forward, orthogonal, reverse and side parking, observation, precedence, correction*	A / B / C	Maneuver
Reversing	*Wrong side, observation, precedence, space, speed, insecure*	A / B / C	Maneuver
Turning	*Place, appropriateness, observation, precedence*	A / B / C	Maneuver
Securing	*Space, sequence, usage of wedge*	A / B / C	Maneuver
Helpers	*Use of persons to support*	A / B	Maneuver
Ramp	*Space, correcting, sideways, backwards*	A / B / C	Maneuver
Parcours	*Lane, slaloming, driving an 8, insufficient, insecure, too fast, aborted, crash*	A / B / C	Maneuver
Emergency brake	*Insufficient, unsafe*	A / B / C	Manuver

Appendix A2

Table A2: Local Assignment

Dependent Variable: *PractExam*

Panel A	I	II	III	IV	V	VI
Diff. in means.	0.035	0.053***	0.043***	0.061***	0.034***	0.032***
p-value	0.153	0.001	0.003	0.000	0.002	0.004
Bandwidth	15-16	14-17	13-18	12-19	11-20	10-21
#L	970	2076	3353	4859	6477	8298
#R	764	1428	1974	2481	2911	3282
Polynomial	1	1	1	1	1	1

Panel B	I	II	III	IV	V	VI
Diff. in means.	0.035	0.053**	0.061***	0.031***	0.076***	0.060***
p-value	0.153	0.019	0.000	0.008	0.000	0.000
Bandwidth	15-16	14-17	13-18	12-19	11-20	10-21
#L	970	2076	3353	4859	6477	8298
#R	764	1428	1974	2481	2911	3282
Polynomial	2	2	2	2	2	2

Notes: *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Bandwidth denotes the window length utilized with *#L* number of observations to the left side of the cutoff and *#R* number of observations to the right including the cutoff. Polynomial denotes the polynomial degree utilized to estimate the regression, which is allowed to differ on both sides of the cutoff.