



**Dishonesty in
healthcare practice: A
behavioral experiment
on upcoding in
neonatology**

Heike Hennig-Schmidt

Department of Economics,
University of Bonn and
Department of Health Management
and Health Economics,
University of Oslo

Hendrik Jürges

Schumpeter School of Business and
Economics, University of
Wuppertal, Germany

Daniel Wiesen

Department of Business
Administration and Health Care
Management, University of
Cologne and Department of Health
Management and Health
Economics, University of Oslo

**UNIVERSITY
OF OSLO**

HEALTH ECONOMICS
RESEARCH NETWORK

Working paper 2018: 3

Dishonesty in healthcare practice: A behavioral experiment on upcoding in neonatology*

Heike Hennig-Schmidt, Hendrik Jürges, Daniel Wiesen

March 25, 2018

Abstract

We introduce a controlled behavioral experiment framed in a neonatal care context to analyze the effect of introducing a random audit and fines on individuals' honesty in a simple reporting task. Our behavioral data provide new evidence on dishonesty and upcoding in health care. We find that introducing audits combined with a fine significantly reduces dishonesty on aggregate. The effect is driven by a significant reduction in upcoding. At the same time, dishonest choices that cannot be detected as fraudulent by an audit (partial dishonesty) increase. We also find evidence that individual characteristics such as gender, medical background, and integrity are related to dishonest behavior.

JEL codes: D03, I11, I18

Keywords: Dishonesty, audits and fines, neonatology, medically framed experiment, reporting of birth weights

1 Introduction

Dishonest behavior significantly affects the cost of medical care provision in many health care systems around the world. One common form of dishonest (or fraudulent) behavior in health

*Hennig-Schmidt: Laboratory for Experimental Economics, Department of Economics, University of Bonn, and Department of Health Management and Health Economics, University of Oslo, Norway; e-mail: hschmidt@uni-bonn.de; Jürges: Schumpeter School of Business and Economics, University of Wuppertal, Germany; e-mail: juerges@wiwi.uni-wuppertal.de; Wiesen: Department of Business Administration and Health Care Management, University of Cologne, and Department of Health Management and Health Economics, University of Oslo, Norway; e-mail: wiesen@wiso.uni-koeln.de. We thank Johannes Abeler, Martin Dufwenberg, Kerstin Eilermann, Gautam Gowrisankaran, Glenn W. Harrison, Keith A. Joiner, Daniele Nosenzo, Charles Noussair, and Gari Walkowitz for their helpful comments and suggestions. We have also benefited from discussions with participants at seminars at the universities of Arizona (Tucson), Cologne, Göttingen, and Wuppertal, as well as at the 2nd Workshop on Behavioral and Experimental Health Economics 2015, Hamilton, Canada, and the 2016 conference of the German Association for Health Economics (dggö) in Berlin, Germany. We thank Emanuel Castillo, Anne Classen, and Lena Kuhne for their excellent programming, research assistance and help in conducting the experiments. Financial support from the University of Wuppertal is gratefully acknowledged. The study was approved by the Ethics Committee of the Medical Faculty of the University of Bonn, Germany.

care is the so-called upcoding of patients to attract higher reimbursements (e.g., Carter et al., 1990; Dafny, 2005). Upcoding typically implies that patients are coded such that their case is related to a more complex illness or treatment, which results in reimbursement for services that are either never provided or not medically necessary.¹

To curtail upcoding, imposing audit mechanisms that include fines for fraudulent behavior is a widely discussed means among healthcare policy-makers.² In general, the classical theoretical motivation for simultaneously implementing audits and fines is the following: dishonest behavior declines whenever the audit probability — assumed to increase the probability of detection—and the size of the fine are such that the expected utility from honesty exceeds the expected utility from dishonesty (e.g., Becker, 1968; Mookherjee and Png, 1989). These considerations emphasize the importance of introducing audits with real monetary consequences for detected fraudulent behaviors.

Empirically, however, it is not well understood how implementing (random) audits including fines affects the health care providers' behavior. Part of the reason is that the use of field data has proven to be quite difficult because of the typical hidden nature of fraudulent behaviors and issues of endogeneity related to introducing audits and fines. Evidence from controlled experiments is also quite limited. In a field experiment, Lindeboom et al. (2016) study the effect of changes in audit rates without a fine on Dutch health care providers' behavior and report almost no evidence for behavioral responses. While this finding suggests the importance of coupling audits with fines (consequences) for fraudulent behaviors, empirical evidence on how audits and fines simultaneously affect behavior is still lacking.

Further, empirical studies on diagnosis-related group (DRG) upcoding typically are based on aggregated (hospital level) data. In neonatology, for example, Jürges and Köberlein (2015) report substantial DRG upcoding at the aggregate level using German birth statistics data. Shi-goeka and Fushimi (2014) analyzed insurance claim data for in-hospital births in Japan. Using

¹These cases constitute improper payments in that the provider has incorrectly (potentially fraudulently) claimed excessive reimbursement on (effective or ineffective) activities that lead to substantial losses for insurers and patients, amounting to billions of dollars. In the financial year 2015, for instance, the improper payment rate which, in addition to fraudulent payment claims, also include lack of documentation and errors for Medicaid was 9.78%, while the Medicare fee-for-service improper payment rate amounted to 12.09% or \$43.33 billion (US Department of Health and Human Services, 2015). Likewise, the German Federal Audit Office (*Bundesrechnungshof*) reported for the year 2010 that one third of hospital billings to incorrectly claimed excessive reimbursement, to the disadvantage of the statutory health insurers; the estimated loss amounted to 875 million Euro out of more than 50 billion Euro of total hospital reimbursements (Bundesrechnungshof, 2011). However, these numbers might conceal even higher rates of potentially undetected fraud (e.g., Becker et al., 2005).

²In the US, for example, a national Health Care Fraud and Abuse Control Program (HCFAC) was established in 1997 to coordinate federal, state and local law enforcement activities with respect to health care fraud and abuse by performing fraud reviews, cost report audits, as well as educating providers and beneficiaries about payment integrity and benefit quality assurance issues. The Office of Inspector General for the US Department of Health and Human Services has the authority to impose civil monetary penalties against providers and suppliers who knowingly submit false claims to the Federal government (Cantrell, 2015). During the fiscal year 2016, over \$3.3 billion was returned to the Federal Government or paid to private persons due to settlements won or negotiated. Since 1997, approximately \$31.0 billion was returned, over \$17.9 billion from 2009 through 2016 (see also US Department of Health and Human Services and Department of Justice, 2017).

these data, it seems difficult to assess how audit mechanisms and fines affect *individual* decision-makers. Moreover, individuals are heterogenous in their behavioral responses to audits and fines and individuals' characteristics such as gender, personalty traits, and integrity might relate to dishonest behavior. Several (neutrally framed) experiments report that women are more honest than men (e.g., Dreber and Johannesson, 2008; Houser et al. 2012) and also personality traits are suggested to influence individuals' behavior in health care contexts (e.g., Donato et al. 2017). However, in health care it is largely unexplored whether and, if so, how individuals' characteristics relate to behavior in fraud-supporting environments (due to, for example, the incentive structure and the absence of audits and fines) and to responses regarding audits and fines.

In this paper, we contribute to filling these gaps. We introduce a controlled behavioral experiment framed in a medical context. The experimental decision situation is motivated by an individual reporting task in neonatal care units. Medical and non-medical students ($N = 98$) decide in the role of obstetrics staff members in a neonatology care unit. The non-linear incentive structure of the experiment mimics the German DRG payments for neonatal care, which primarily depend on the birth weight from the first weighing recorded in the birth report. We consider two thresholds (1,250g and 1,500g) at which the DRG payment almost doubles when reporting a birth weight below the respective thresholds. Subjects receive a lump-sum payment and a variable component depending on their birth weight reports. Costs for medical treatment depend on the true birth weight and are deducted from the DRG payment. Therefore, subjects can increase profits significantly when they upcode by fraudulently reporting a birth weight lower than the DRG thresholds (due to the non-linear, discrete DRG scheme); we call such behavior dishonest.

In addition to a baseline No-Audit treatment, in which upcoding cannot be detected, we introduce, in a between-subject design, a treatment with random audits and fines. Experimental parameters are such that only discrepancies of more than 50g between the reported birth weight and the actual birth weight can be detected by an audit. Smaller deviations cannot be discovered. Audits take place with a 10% probability. Detected dishonesty implies a fine, amounting to discarding all the aggregate earnings for the reporting task. In the baseline treatment without audit, dishonesty has no financial consequences for subjects. Further, since we can identify cheating individually, our experimental design allows us to study the link between dishonesty and personality traits such as the Big Five (Rammstedt and John, 2007, Gosling et al., 2003) or integrity (Schlenker, 2008).

In this setting, we address the following research questions. Research question 1: Do subjects in the baseline treatment engage in DRG upcoding (i.e., dishonest behavior), and if so, to what extent? Research question 2: Does introducing audits and fines reduce the upcoding of birth weights? If so, to what extent is upcoding reduced? Research question 3: Does introducing audits and fines lead to less upcoding in general or only to less upcoding at birth weights where dishonesty becomes detectable through audits? Research question 4: How do the individual

characteristics (a) major in medicine or economics and (b) gender relate to upcoding and how do these relate to behavioral responses when audits and fines are introduced? Research question 5: How do individuals' (a) personality traits, (b) risk aversion, and (c) integrity relate to dishonesty and to behavioral responses to audits and fines?

Our behavioral data evidences substantial upcoding. We find that nearly 75% of all decisions in the No-Audit treatment are dishonest, in that they increase subjects' payoffs by upcoding. When excluding birth weights where there is no financial gain from dishonesty, the proportion of dishonest reports is 87%. When participants are given the choice of partial or full dishonesty, 70% of the participants are fully dishonest and 25% are partially dishonest. Only 5% are honest. Introducing audits and fines significantly reduces upcoding by 27 percentage points. This reduction is concentrated on birth weights where detection is possible. At birth weights where participants can choose to be partially dishonest, honest entries increase by 15%. In addition, we find a shift from (detectable) fully dishonest to (non-detectable) partially dishonest behavior: the latter doubles while the former is cut by half. Participants seem to prefer to choose the safe but lower profit rather than the risky but maximal return.

Individual characteristics are related to upcoding and responses to Audit-and-fine. We find that medical students are more dishonest under No-Audit than economics students, but introducing Audit-and-fine affects medical students more strongly than economics students. Men and women are similarly dishonest under No-Audit, but women react more strongly to Audit-and-fine than men. Personality traits do not relate to upcoding behavior in a robust and systematic way. Further, we find that risk-averse individuals tend to report birth weights less fraudulently, although not significantly so. The integrity of individuals is also related to upcoding: individuals with higher integrity report birth weights more honestly than individuals with lower integrity.

The paper proceeds as follows. In Section 2, we relate our study to the literature. Section 3 reports the institutional background of neonatal care units in Germany, and Section 4 describes the experimental design and procedure. In Section 5, we present the results from our experiment. Section 6 discusses results of our study and Section 7 concludes.

2 Related literature

Our study contributes to several streams of the literature. First, we complement empirical research on DRG upcoding in health care markets. Early studies have already pointed to the unintended consequences of introducing DRG-based payment systems,³ namely the attempt by hospitals to increase revenues through fraudulent coding practices (e.g., Carter et al., 1990; Silverman and Skinner, 2004; Dafny, 2005; Januleviciute et al., 2016; Barros and Braun, 2017).

³Prospective payment systems using diagnosis-related groups (DRGs) to reimburse hospital inpatient care became very popular two decades ago (Ellis and McGuire, 1993). The main reason for this popularity was that they are thought to have predominantly desirable effects on hospital efficiency and quality.

Shigeoka and Fushimi (2014), Jürges and Köberlein (2015), and Reif et al. (2018) report evidence for DRG upcoding in neonatal care in Japan and Germany. Common to all these studies is that evidence is provided at an aggregate level. We complement the latter studies in particular by analyzing upcoding at an individual subject level in a neonatal care context.

Second, we relate to the literature on audits. Empirical evidence on the effectiveness of audits is rather scarce and primarily comes from neutrally framed laboratory experiments. For example, Clark et al. (2004) analyze random auditing, a conditional audit rule and an optimal audit rule. They find the latter two to be associated with fewer audits, whereas the compliance rate is maximized under random auditing. Cason and Gangadharan (2006) propose an experiment addressing Harrington’s (1988) paradox (compliance to regulations even though the frequency of inspections and audits is low), in which they vary the probability of inspection and the severity of a fine. They report that compliance behavior does not change as sharply as the model predicts. In a field experiment, Lindeboom et al. (2016) study the effect of changes in audit rates *without* a fine on the behavior of Dutch health care providers, reporting almost no evidence for behavioral responses. Our medically framed experiment allows us to compare two situations: (i) No audit and (ii) Audit and fine, a binding, random audit rule (in 10% of the cases an audit takes place) and a considerable fine if dishonest reporting is detected.⁴ The fine implies that the variable payment for the birth weight reporting task is foregone when fraudulent behavior is detected in at least one of the six entries. We thus provide further evidence on the effect of auditing on the behavior of individuals in a health care setting.

Third, we relate to the literature linking individual characteristics and (dishonest) behavior. On the one hand, evidence on the impact of gender on dishonest behavior is rather mixed. Abeler et al. (2014) report no gender differences in dishonest behavior for a representative sample of subjects. Childs (2012) and Gylfason et al. (2013) find no gender differences either. However, there is considerable experimental evidence that women are more honest than men (e.g., Dreber and Johannesson, 2008; Houser et al. 2012; Muehlheusser et al., 2015). Further, personality traits are often suggested as influencing the behavior of individuals (e.g., Ben-Ner et al., 2004). In a health care setting, Donato et al. (2017) report that personality traits correlate with individuals’ responses to incentive contracts. To the best of our knowledge, empirical evidence on the relationship between personality traits and dishonest behavior in a medical context is lacking. We employ a 10-item short-version questionnaire of the Big Five Inventory (Gosling et al., 2003, Rammstedt and John, 2007) to elicit personality traits and relate them to dishonesty. Moreover, we link the subjects’ reporting behavior to their medical major (Hennig-Schmidt and Wiesen, 2014), to integrity (Schlenker, 2008), and to their risk attitudes measured by a simple questionnaire item (Dohmen et al., 2011). While adding to the literature by exploring the impact of sociodemographics on dishonest behavior, we further explore the link between these

⁴In a theoretical model, Kuhn and Siciliani (2013) differentiate binding and non-binding commitment to audit when deriving optimal pricing and auditing policies from the purchaser’s perspective.

individual characteristics and upcoding in a neonatal care setting. We thus complement the empirical literature on DRG upcoding, which, due to aggregate data, is not able to analyze how individual characteristics relate to dishonest behavior.

Finally, our study is also related to the growing experimental economics literature analyzing dishonest behavior; see Abeler et al. (2016) for a comprehensive review. Dishonest behavior is analyzed, for example, in the ‘die-in-a-cup’ game (Fischbacher and Föllmi-Heusi, 2013), a coin toss (Abeler et al., 2014), or a matrix task (Mazar et al., 2008). The key difference of our study compared to these earlier papers is not only the medical frame, in particular the neonatology context, of the experiment. It is just this frame that seems to provide arguments for justifying dishonest behavior. Such frame-induced motives are absent in the above studies.⁵ We will come back to this issue in the discussion section.

3 Institutional background: Neonatal care units in Germany

We now present some background on neonatal care units in Germany, and we explain the DRG reimbursement scheme and the embedded incentives for upcoding. Since 2003, reimbursement for neonatal care in German hospitals is based on the following case characteristics: birth weight, surgical (OR-) procedures, long-term artificial respiration, complications, and 5-day and 28-day mortality. Birth weight is the most important determinant: DRGs are defined along eight birth weight thresholds: 600g, 750g, 875g, 1,000g, 1,250g, 1,500g, 2,000g, and 2,500g. Reimbursement substantially rises at each threshold, i.e., very small differences in birth weight can yield differences in reimbursements by 15,000 EUR (about US-\$ 17,800) or more. Since reimbursement for each DRG is determined to cover the true average treatment costs in that group, all newborns with a birth weight just above some threshold value (if reported honestly) will create a financial loss to the hospital, whereas all newborns with a true birth weight just below some threshold will create financial gain.

As documented in Jürges and Köberlein (2015) and Reif et al. (2018), hospitals appear to react by manipulating the recorded birth weight and by re-classifying newborns to better-paying DRGs. While the true treatment costs of each newborn remain the same, reimbursement will jump from lower-than-expected treatment costs to higher-than-expected treatment costs. The current setting embeds a strong incentive for dishonest reporting. This incentive is reinforced by the fact that individual cases of birth weight manipulation (within reasonable limits) can

⁵Note that we are among the first to analyze dishonest behavior by individual decision data in a procedure that allows us to observe whether an individual report is truthful or not. Yet, individual decisions cannot be traced back to the individual subject. To the best of our knowledge, only Abeler et al. (2016) and Gneezy et al. (2018) use a similar procedure as we do. In none of the other previous experimental studies can an individual report be identified as truthful. The researcher, however, can judge whether the reports are biased towards better-paid outcomes by comparing behavioral data with a random distribution. Our paper further contributes to the literature by using an experimental design that — unlike most ‘die-in-a-cup’-game studies — provides constant payoffs for several different weights (rolled numbers in the ‘die-in-a-cup’ design).

hardly be detected ex post. Newborn infants lose around 4 to 7% of their weight within the first 24 hours due to initial postnatal weight loss (e.g., Flaherman et al., 2015), which makes detection even more difficult. Since DRGs in neonatology are partly determined by the reported birth weight entered in the birth report, obstetrics on the ward are actually responsible for the “coding” and not, as one might assume, a coder in the hospital’s administration.

DRG upcoding in neonatology leads to substantial financial losses for health insurers. Since the introduction of DRGs, hospitals have gained additional reimbursement in excess of 100 million Euro (Jürges and Köberlein, 2015). Thus, research needs to focus on the determinants of DRG upcoding in neonatology to understand better the decision-makers’ underlying motivations. This would enable policy-makers to design reimbursement systems, but also institutions that reduce the incentives for fraudulent behavior like, for instance, (tighter) audits and fines often debated among health care policy-makers.

4 The experiment

4.1 Design

In our medically framed experiment, subjects take the roles of obstetrics staff members. Each subject i is entrusted with recording the birth weight (w_j) of six preterm infants j into their neonatal birth reports. Reported birth weights can take nine different values in grams $w_j \in [1, 150, \dots, 1, 550]$ in intervals of 50 grams. Subject i decides which birth weight (\hat{w}_j^i) to enter into the birth record after having seen the true weight w_j on the computer screen. Six true birth weights are randomly drawn and shown in random order on the subjects’ screens: these are 1,200, 1,250, 1,300, 1,350, 1,400, and 1,500 grams. These birth weights are the same for all subjects, and each birth weight occurs only once.

Subjects receive a fixed lump-sum payment F and a variable payment depending on the reported birth weights. The individually reported birth weights \hat{w}_j^i determine subject i ’s payment (DRG reimbursement rates) per case $p(\hat{w}_j^i)$. The medical treatment of the preterm infants incurs costs c , which are, on average, almost compensated by the lump-sum reimbursement. The costs $c(w_j)$ depend on the true birth weight, w_j , displayed on the scale. All six decisions are payoff-relevant. Subject i ’s overall profit is therefore $\pi^i(w_j, \hat{w}_j) = F + \sum_{j=1}^6 p(\hat{w}_j) - c(w_j)$. Further, we assume that each early-born infant receives optimal medical care according to its true birth weight w_j — irrespective of the reported birth weight \hat{w}_j^i .⁶

Figure 1 illustrates the DRG reimbursement rates and costs in our experiment. We choose birth weights to be in the middle range of real-world figures for preterm infants, which allow us to consider two thresholds (one at 1,250g and another at 1,500g). The parameter values for the

⁶This assumption seems reasonable, in light of recent empirical evidence by Reif et al. (2018), which indicates that financial incentives related to the birth weight thresholds do not directly impact the medical care that newborns receive.

Figure 1: DRG reimbursement rates and medical treatment cost



Notes: This figure shows the DRG reimbursement rates and the average medical treatment costs associated with the reported and the true birth weights, respectively. The birth weight thresholds are adopted from the German DRG system and the values for reimbursement and cost are indicated in Taler, our experimental currency. These values are calibrated using actual DRG reimbursement rates as reference values.

DRG reimbursement rates in our experiment are adapted to the German DRG system. They are set such that the average costs for the medical treatment of an early-born infant are almost covered within the DRGs. Table 1 shows the profit levels for each combination of reported and true birth weight.

Reporting lower than true birth weights *within* a DRG has no consequences for the subjects' profits, as the reimbursement is fixed within the same DRG and the cost to be deducted is based on the true birth weight. It is only by upcoding that subjects gain additional profit—in particular, when subjects report a birth weight lower than the next or next but one lower threshold.

At the true birth weight of 1,500g we can distinguish between full and partial dishonesty. Subjects can either upcode by one or by two DRG thresholds. We thus are able to analyze whether subjects – in order to receive the maximum profit – upcode to the full extent (i.e., report $\hat{w}_j = 1,200$ or below). We define this reporting behavior as *fully dishonest*. When upcoding by one DRG threshold (i.e., reporting $1,250 \leq \hat{w}_j < 1,450$), subjects receive a payment higher than by reporting honestly, but lower than the maximum profit. We define this behavior as *partially dishonest*. For all birth weights, we call reporting behavior *honest* when subjects enter the true

Table 1: Profit matrix in the experiment

True birth weight w_j (in g)	Reported birth weight \hat{w}_j (in g)								
	1150	1200	1250	1300	1350	1400	1450	1500	1550
1200	80	80	-100	-100	-100	-100	-100	-180	-180
1250	120	120	-60	-60	-60	-60	-60	-140	-140
1300	150	150	-30	-30	-30	-30	-30	-110	-110
1350	180	180	0	0	0	0	0	-80	-80
1400	200	200	20	20	20	20	20	-60	-60
1500	240	240	60	60	60	60	60	-20	-20

Notes: This table shows profits for the weights in the subjects' choice range (first column) and the birth weights that can be entered in the birth report (second row). Note that reimbursements depend on the reported birth weights. DRG thresholds are at 1,250 and 1,500. Costs depend on the true birth weight and increase with decreasing birth weights.

birth weight shown on their computer screen. Recall that partial and full dishonesty have been indicated in the experimental literature as full and partial lying, respectively (e.g., Fischbacher and Föllmi-Heusi, 2013; Abeler et al., 2014, 2016).⁷

4.2 Treatments

At a between-subject level, we vary the detection probability of fraudulently reported birth weights. In the baseline treatment No-Audit, the detection probability is zero, and the subjects' upcoding has no financial or other consequences. In treatment Audit-and-fine, the subjects' reported birth weights are audited with a probability of 10%. In case a subject is audited and at least one misreported birth weight is detected, all earnings based on the DRG reimbursement rates are discarded. Subjects then are paid out the fixed amount F only.

Our audit mechanism makes use of the fact that a newborn baby loses a certain percentage of its birth weight in the first 24 hours—we use 5% in our experiment (see the Instructions in Appendix A.1). If the weight measured on the day after birth is higher than the recorded birth weight, the latter is too low, which constitutes attempted fraud. After all decisions have been taken, the subjects are informed whether the birth weights that subjects had entered were audited and whether a false entry was discovered. Table 2 provides an overview on our experimental treatments.

One distinct feature of our experiment is that, even in the Audit-and-fine treatment, participants can behave dishonestly without fear of being detected at some of the birth weights. Due to our parameter specification, underreporting of birth weights up to 50g *cannot* be detected, while larger manipulations within and between DRGs can. This feature allows a comparison of

⁷On explanations for the aversion of individuals to lie, see, for example, Kartik (2009), Dufwenberg and Dufwenberg (2018) and Gneezy et al. (2018).

Table 2: Experimental treatments

Treatment	Detection probability	Payment	Number of subjects (medical students)
Baseline No-Audit (no audit and no fine)	0%	$F + \sum_{j=1}^6 p(\hat{w}_j) - c(w_j)$	44 (27)
Audit-and-fine	10%	<ul style="list-style-type: none"> ◦ F, if cheating is detected ◦ $F + \sum_{j=1}^6 p(\hat{w}_j) - c(w_j)$, otherwise 	54 (24)

Notes: This table shows treatments, the parameters for detection probabilities, the payment formula, and the number of participants (medical students).

upcoding with and without detection probability in the same treatment. Specifically, upcoding is attractive at true birth weights just above remuneration thresholds, as a comparatively small manipulation by 50g will increase profits at no detection risk. For true birth weights just above the DRG thresholds (1,250g and 1,500g) subjects can upcode without the threat of being detected.⁸ Recall the rather serious fine of discarding all profits from entering the six birth weights for a participating subject if one of his/her reports is found to understate the birth weight by more than 50g.

4.3 Protocol

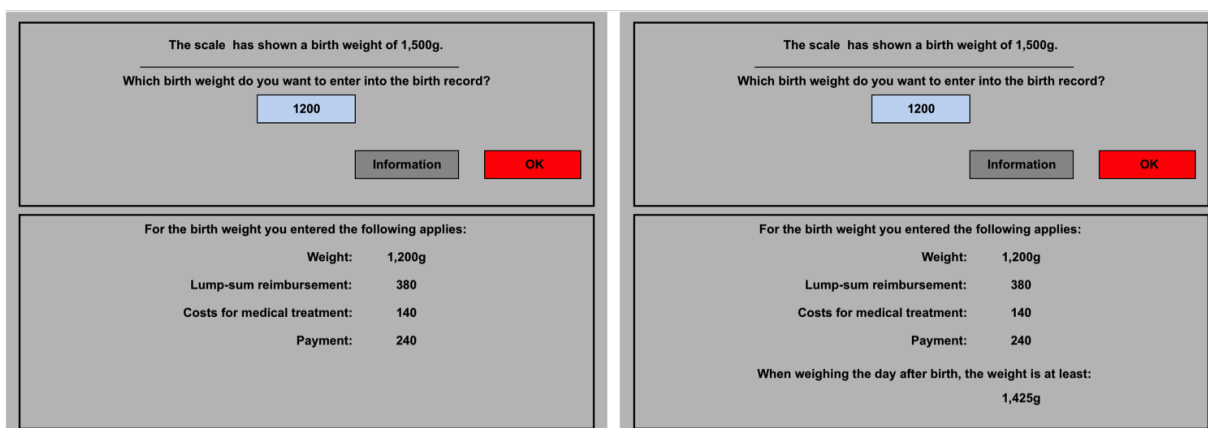
The computerized experiment was programmed with z-Tree (Fischbacher, 2007) and conducted at *BonnEconLab*, the Laboratory for Experimental Economics at the University of Bonn, in May and December 2014 and in December 2015. Overall, 98 students participated in our experimental sessions. Among those were 51 medical students, of whom 27 (24) participated in the No-Audit (Audit-and-fine) treatment. Of the 47 economics students, 17 took part in the No-Audit and 30 in the Audit-and-fine condition. The average age was 23 years.

Subjects were recruited via the online recruiting systems ORSEE (Greiner, 2015) and hroot (Bock et al., 2014). The recruiting process of subjects was the following: Students who registered in ORSEE for laboratory experiments at *BonnEconLab* were invited via automatically generated e-mails to participate in the experimental conditions of our experiment. When signing up, subjects did not know about the decision task, the composition of subjects, or about our research objective. This procedure guaranteed the random allocation of students to experimental treatments and excluded self-selection into audit conditions.

The procedure in both treatments was as follows: Upon arrival, subjects were randomly

⁸For example, when $w_j = 1,250\text{g}$, reporting a birth weight of $\hat{w}_j = 1,200\text{g}$ cannot be detected by an audit, as the reported birth weight is higher than the lower bound of the true weight of 1,187.5g at the second weighing. Another example is $w_j = 1,500\text{g}$ (lower bound of the true weight at the second weighing: 1,425g). Here, reporting $\hat{w}_j = 1,450\text{g}$ (partial dishonesty) cannot be detected by an audit, whereas reporting $w_j = 1,200\text{g}$ (full dishonesty)

Figure 2: Screen shots for treatments “No-Audit” and “Audit-and-fine”



Notes. The left panel shows a sample decision screen in the baseline treatment and the right panel shows a screen from the Audit-and-fine treatment. In both screens, subjects are asked to enter the birth weight, which is displayed on top of the screen (here: 1,500 grams). Subjects can receive additional information when clicking on the button labeled “Information”. They are informed on the reimbursement based on their entry, the costs based on the true birth weight, and the resulting profit. In the Audit-and-fine condition, subjects are provided with additional information on the lower bound of the newborn’s weight at the second weighing (here: 1,425 grams).

allocated to cubicles. They were then given ample time to read the instructions and ask clarifying questions, which were answered in private. To verify that the subjects had understood the decision task, they had to answer a set of control questions. The experiment did not start unless all subjects had answered all control questions correctly.⁹ Then, subjects anonymously decided for six early-born infants, which birth weight to enter into each early-born’s medical record after having seen the true birth weight on the computer screen. Figure 2 provides an illustration of the screens in both treatments. The birth weights subjects faced were given in individualized pre-determined random order. The introduction stressed that the subjects’ decisions were confidential and anonymous and that the experimenter could not identify who had entered which birth weight.

Before entering the birth weight of a specific baby, subjects could learn about the respective lump-sum remuneration, the treatment cost for each birth weight, and the resulting profit by clicking an information button on their computer screen. In the Audit-and-fine treatment, subjects could also learn about the lower bound of the preterm infant’s weight on the day after birth. All monetary amounts were given in Taler, our experimental currency, the exchange rate

can.

⁹For instructions and control questions, see Appendix A.1

being 1 Taler = 0.01 EUR.¹⁰ Subjects were also informed about the fixed payment F amounting to 400 Taler.

After having made all six decisions, subjects were asked to answer a questionnaire on gender, age, attitudes, and personality. Subjects received an additional four Euro for answering the questionnaire. Finally, they were paid by a confidential payment procedure that did not allow us to trace any individual subjects' decisions, and they subsequently left the laboratory. Sessions lasted about 45 minutes, including filling in the post-experimental questionnaire. Subjects earned 10.30 Euro on average, including the payment of 4 Euro for answering the questionnaires (Audit: 11.86 Euro, Audit-and-fine: 9.02 Euro).

5 Results

5.1 Reporting behavior in treatment “No-Audit”

We first address Research question 1 which is concerned with the degree of the subjects' dishonesty in the baseline “No Audit” treatment. We further investigate whether subjects upcode to the full extent or only partially, i.e., we analyze whether for the true birth weight of 1,500g subjects upcode by two thresholds (by reporting birth weights even below 1,250g) or whether they upcode to the next DRG threshold only and forgo the maximum profit (partial dishonesty).

We start with a descriptive analysis of the full data set of 264 entries by showing the distribution of true and reported birth weights under the baseline No-Audit treatment. Table 3 provides an overview on the subjects' reported birth weights for all true birth weights. Light gray cells indicate underreported birth weights. Darker cells show correctly reported weights, while white cells denote overstated birth weights. Table 3 indicates that a substantial share of reported birth weights is below the true weight (light gray shaded area) and a strong cumulation below the 1,250g threshold.

In Table 4, we further categorize the subjects' reporting behavior. Only 22% of all birth-weight entries under No-Audit are honest (57 of the 264 decisions). The majority of reported birth weights imply upcoding, namely 73% (192 decisions). That means nearly three quarters of all decisions in the No-Audit condition increase the subjects' payoffs by fraudulent upcoding. 61% of these (162 decisions) comprise upcoding by one, 11% (30 decisions) by two DRGs. The remaining 6% of decisions imply reporting birth weights that are not payoff increasing — either within a DRG or above a DRG threshold (which implies higher cost and therefore lower payoffs compared to honest reporting; recall Table 1). When looking only at the 220 decisions where upcoding is possible (excluding the true weight of 1,200g) and where payoffs can actively be

¹⁰The exchange rate was calculated to result in an hourly wage of a student assistant at the University of Bonn of around 10 Euro.

Table 3: Distribution of reported birth weights by true birth weights: No-Audit treatment

True weight	Reported weight									Total
	1150	1200	1250	1300	1350	1400	1450	1500	1550	
1200 (abs.)	4	37	2	0	0	0	0	1	0	44
1200 (%)	9	84	5	0	0	0	0	2	0	100
1250	4	37	2	0	0	1	0	0	0	44
	9	84	5	0	0	2	0	0	0	100
1300	6	32	2	4	0	0	0	0	0	44
	14	73	5	9	0	0	0	0	0	100
1350	7	28	0	1	7	0	0	0	1	44
	16	64	0	2	16	0	0	0	2	100
1400	5	31	0	1	1	6	0	0	0	44
	11	70	0	2	2	14	0	0	0	100
1500	4	26	0	0	0	1	11	1	1	44
	9	59	0	0	0	2	25	2	2	100
Total	30	191	6	6	8	8	11	2	2	264
	11	72	2	2	3	3	4	1	1	100

Note: Light gray cells indicate understated birth weights. Darker cells show correctly reported birth weights, while white cells denote overstated birth weights. Numbers in the first (second) line for each true weight indicate absolute frequencies (percentages). Data set for the No-Audit treatment consists of 264 entries ($N = 44$ subjects).

increased by fraudulent reporting, we find 87% dishonest answers.

At the true birth weight of 1,500g, only one person (2%) decides honestly, 27% (12 out of 44 subjects) upcode by one threshold and behave partially dishonestly, which means they make less than the maximum profit; 68% (30 of the 44 participants) make the maximum profit by upcoding across two thresholds, thus behaving fully dishonestly; see Table 3. In sum, we state:

Result 1. *In the No-Audit treatment, nearly three quarters of all decisions imply upcoding (dishonest report of birth weight, which increases an individual’s payoff). For the true birth weight of 1,500g (at which different levels of dishonest behavior are distinguishable), about 70% of all decisions are fully dishonest, about 25% are partially dishonest, and 5% behave honestly.*

5.2 Reporting behavior in treatment “Audit-and-fine”

We now investigate whether the introduction of a random audit (10% detection probability) coupled with a substantial fine reduces upcoding (Research question 2). Table 5 shows the frequency of reported weights for the true birth weights in the Audit-and-fine treatment.

We find that introducing random audit and a fine substantially affects the subjects’ reporting behavior. The percentage of all honest birth-weight reports almost doubles to 43% (140 of the 324 decisions); see Table 4. This is a highly significant increase by 22 percentage points compared to the baseline treatment ($p \leq 0.001$, Fisher exact test, two-tailed). Upcoding is reduced to

Table 4: Reporting behavior by experimental treatments

	No-Audit		Audit-and-fine	
	Abs. freq.	Percent	Abs. freq.	Percent
Honest reporting	57	21.6	140	43.2
Upcoding: dishonest reporting (payoff-increasing)	192	72.7	162	50.0
One DRG threshold	162	61.4	144	44.44
Reported weight $\hat{w}_j^i - w_j \leq 50\text{g}$ / non-detectable	(48)	(18.2)	72	22.2
Reported weight $\hat{w}_j^i - w_j > 50\text{g}$ / detectable	(114)	(43.2)	72	22.2
Two DRG thresholds / detectable	30	11.4	18	5.6
Other misreporting (not payoff-increasing)	15	5.7	22	6.8
Total	264	100.0	324	100.0

Notes: This table provides an overview of the subjects' reporting behavior. In category 'Other misreporting', we consider reported birth weights that imply no financial gain and comprise underreported true birth weights within the same DRG, as well as over-reported birth weights. Numbers in parentheses refer to entries in No-Audit at those weights where upcoding is (non)detectable in Audit-and-fine.

50%, while other misreporting comprises 7% of all choices. When analyzing decisions only where upcoding is possible (excluding entries for $w_j = 1,200\text{g}$), 60% of birth weights are reported dishonestly; a reduction by around 27 percentage points.

Estimation results from logit models confirm the non-parametric analyses. Introducing audits and fines significantly reduces the likelihood of upcoding; see Model (1) in Table 6. This finding is robust when controlling for individuals' gender, major, and other individuals' characteristics; see Model (2) in Table 6. In sum, we state the following result:

Result 2. *On aggregate, introducing a random audit (10% detection probability) coupled with a fine significantly reduces dishonest behavior. Honest reports increase, while upcoding is significantly reduced.*

Next, we analyze whether introducing audits and fines leads to more honest reporting in general or whether we observe less upcoding only at birth weights where dishonesty is detectable (Research question 3). Estimation results show that upcoding is only reduced significantly at birth weights where upcoding can be detected; see Model (3) in Table 6. The coefficient of the interaction term of 'Audit-and-fine \times detection' indicates a decrease of about 38 percentage points. The main effect is robust when including individual controls; see Model (4) in Table 6. At birth weights where detection is not possible (1,250g) the effect of audit is small and statistically insignificant. Upcoding does significantly differ between birth weights where detection is possible and where it is not; see χ^2 -values for Models (3) and (4) in Table 6. In sum, our treatment variation makes an important point concerning audits and fines. They appear not to

Table 5: Frequency of reported birth weights by true birth weights: Audit-and-fine treatment

True birth weight	Reported birth weight									Total
	1150	1200	1250	1300	1350	1400	1450	1500	1550	
1200 (abs.)	4	49	0	0	0	0	0	0	1	54
(%)	7	91	0	0	0	0	0	0	2	100
1250	5	45	4	0	0	0	0	0	0	54
	9	83	7	0	0	0	0	0	0	100
1300	6	23	4	19	2	0	0	0	0	54
	11	43	7	35	4	0	0	0	0	100
1350	7	12	0	3	29	2	1	0	0	54
	13	22	0	6	54	4	2	0	0	100
1400	6	13	0	0	2	30	3	0	0	54
	11	24	0	0	4	56	6	0	0	100
1500	7	11	0	0	0	0	27	9	0	54
	13	20	0	0	0	0	50	17	0	100
Total	35	153	8	22	33	32	31	9	1	324
	11	47	2	7	10	10	10	3	0	100

Note: Light gray cells indicate understated birth weight that could be detected by audit. Medium gray cells indicate understated birth weight that could not be detected by audit. Dark gray cells indicate correctly stated birth weight, whereas white cells denote overstated birth weights. Numbers in the first (second) line for each true weight indicate absolute frequencies (percentages). $N = 54$, data set of 324 entries for treatment Audit-and-fine.

cause more honest behavior in general, but reduce the willingness to engage in upcoding only where it is detectable.

Finally, we analyze behavior at the true birth weight of 1,500g. Table 5 shows that honest behavior rises to 17% (9 of 54 decisions) — compared to 2% under No-Audit. Upcoding that cannot be detected by an audit (reporting 1,450g instead of 1,500g) rises to 50%. This is a significant increase of 22 percentage points compared to the baseline treatment ($p = 0.013$, Fisher exact test, two-tailed). Estimation results from a multinomial logit model on the probability of being honest, partially dishonest, and fully dishonest confirm and extend the non-parametric analysis; see Table 7. Model (1) shows that introducing Audit-and-fine reduces the probability of being fully dishonest significantly by 40 percentage points. This reduction is accompanied by a strong shift from fully dishonest to partially dishonest behavior and by a smaller percentage-point increase of honest entries. Many participants thus chose the safe, but lower profit of 60 Taler, rather than the risky but higher return of 240 Taler (recall Table 1). This finding is robust when including covariates; see Model (2) in Table 7. In sum, we state the following result:

Result 3. *Introducing a random audit (10% detection probability) coupled with a fine significantly reduces the likelihood of upcoding only if fraudulent reporting is detectable. At the birth weight of 1,500g, fully dishonest entries, which are detectable, decrease significantly while honest and partially dishonest choices, which are non-detectable, significantly rise.*

Table 6: Effect of Audit-and-fine on the probability of upcoding

	Dependent variable: upcoding			
	(1)	(2)	(3)	(4)
Mean dep. var (No-Audit)	0.873	0.873	0.852	0.852
Audit-and-fine	-0.273** (0.052)	-0.236** (0.054)		
Audit-and-fine \times detection			-0.379** (0.075)	-0.323** (0.074)
Audit-and-fine \times no detection			-0.015 (0.130)	0.049 (0.120)
Male		0.089* (0.051)		0.111* (0.061)
Medicine		0.000 (0.062)		0.028 (0.071)
χ^2 detect vs. no detect			4.852**	6.556**
Birth weight FE	Yes	Yes	Yes	Yes
Individual controls	No	Yes	No	Yes
N decisions	490	490	392	392
N subjects	98	98	98	98

Notes: This table reports average marginal effects based on logit models. Standard errors are clustered at the individual subject level shown in parentheses. We consider decisions only where upcoding is possible, which means we exclude entries for $w_j = 1, 200g$. This results in a data set of 490 decisions. The dependent variable ‘Upcoding’ = 1 if individuals misreport birth weights such that their payoff increases, and = 0 otherwise. Individual control variables comprise four age categories (18-20, 21-25, 25-29, 30+), personality traits (Big Five inventory), and integrity. In Models (3) and (4), the data for 1,500g are excluded due to the detection possibility being endogenous. * $p < 0.1$; ** $p < 0.05$.

5.3 Students’ major and gender

We now analyze how individuals’ major and gender relate to dishonest behavior and to behavioral responses when audits and fines are introduced (Research question 4). Table 8 shows that, in the absence of audit and fine, medical students are on average more likely to upcode compared to economists (by 10 percentage points). Introducing audits has a significant effect only on medical students’ upcoding, reducing upcoding by 43 percentage points. Economics students, on the other hand, show an insignificant response to the introduction of audit and fine; see Models (1) and (2) in Table 8. The difference between medical and economics students in responding to audit and fine is significant at the five-percent level; see z -value difference for estimates of Models (1) and (2).

When separating between birth weights where detection is possible and where detection is not possible, we find further differences between medical and economics students. The likelihood of upcoding is reduced by nearly 60 percentage points for medical students, but only by 10

Table 7: Effect of Audit-and-fine on the probability of being honest, partially dishonest, and fully dishonest at true birth weight 1,500g

	Without covariates (1)			With covariates (2)		
	Honest	Partially Dishonest	Fully Dishonest	Honest	Partially Dishonest	Fully Dishonest
Mean dep. var. (No-Audit)	0.045	0.250	0.705	0.045	0.250	0.705
Audit-and-fine	0.144** (0.060)	0.261** (0.094)	-0.406** (0.090)	0.122* (0.070)	0.202* (0.106)	-0.325** (0.099)
Medical student	0.157** (0.062)	0.054 (0.095)	-0.211** (0.090)	0.183** (0.085)	-0.072 (0.108)	-0.111 (0.098)
N decisions		98			98	
N subjects		98			98	

Notes: This table reports average marginal effects based on multinomial logit models. Standard errors are shown in parentheses. Individual control variables in Model (2) comprise four age categories (18-20, 21-25, 25-29, 30+), personality traits (Big Five inventory), and integrity. * $p < 0.1$; ** $p < 0.05$.

percentage points for economists. At birth weights where detection is not possible, however, the likelihood of upcoding increases significantly for economists by about 30 percentage points compared to No-Audit; see Models (5) and (6) in Table 8. Medical students, in contrast, reduce upcoding, albeit not significantly. Again, we find a significant difference between medical and economics students; see z -value difference for estimates of Models (5) and (6).

As to gender, in treatment “No audit”, upcoding does not differ between females and males. Introducing audit and fine significantly reduces the likelihood for upcoding for both genders. The magnitude of the response is higher for women (36 percentage points) than for men (16 percentage points). This difference is weakly significant; see Models (3) and (4) in Table 8; see also Models (2) and (4) in Table 6. According to Models (7) and (8), we observe a similar difference between female and male subjects when detection is possible. When upcoding cannot be detected, we find no significant response compared to the No-Audit treatment. The differences between weights where detection is possible and where it is not are always significant at the five-percent level, for major as well as for gender, see Models (5) to (8). In sum, we state:

Result 4. (a) *In the No-Audit treatment, medical students behave slightly more dishonestly than economics students. When a random audit with a fine is introduced, medical students reduce their upcoding significantly, while economists do not. Where detection is possible, medical students significantly reduce upcoding. Where detection is not possible, economists increase their upcoding significantly, which is in contrast to medical students.* (b) *Upcoding is similar between males and females under No-Audit. The reduction in upcoding under Audit-and-fine for females is significantly larger than for males; the same applies when detection is possible. If detection is not possible, no gender difference is observed.*

Table 8: Effect of Audit-and-fine on the probability of upcoding by gender and field of study

	Dependent variable: upcoding							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Econ	Med	Women	Men	Econ	Med	Women	Men
Mean dep. var (no audit)	0.812	0.911	0.867	0.882	0.765	0.907	0.843	0.868
Audit-and-fine	-0.063 (0.076)	-0.429** (0.065)	-0.363** (0.072)	-0.157** (0.067)				
<i>z</i> -value difference ^a	1.86*							
Audit-and-fine × detection					-0.110 (0.078)	-0.593** (0.080)	-0.502** (0.100)	-0.244** (0.076)
<i>z</i> -value difference	4.42**							
Audit-and-fine × no detection					0.302** (0.050)	-0.118 (0.077)	0.044 (0.092)	0.110 (0.079)
<i>z</i> -value difference	3.32**							
χ^2 detect v no detect					5.90**	8.14**	11.07**	6.04**
Birth weight FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>N</i> decisions	235	255	260	230	188	204	208	184
<i>N</i> subjects	47	51	52	46	47	51	52	46

Notes: This table shows average marginal effects based on logit models. Standard errors are clustered at the individual subject level shown in parentheses. We consider decisions only where upcoding is possible which means we exclude entries for $w_j = 1, 200g$. The dependent variable 'Upcoding' = 1 if individuals misreport birth weights such that their payoff increases, and = 0 otherwise. Individual control variables: four age categories (18-20, 21-25, 25-29, 30+), gender (in columns 1, 2, 5, 6), field of study (columns 3, 4, 7, 8), Big Five Inventory, and integrity. In Models (5) to (8), the data for 1,500g are excluded due to the detection possibility being endogenous. * $p < 0.1$; ** $p < 0.05$. ^a*z*-values for difference of estimated parameters *across* equations were obtained as *z*-values of interaction effects in fully interacted models.

5.4 Personality traits, integrity, and risk aversion

In this section, we address Research question 5 on how the individuals’ personality traits, integrity, and risk aversion relate to upcoding and responses to audits and fines. To this end, we use the short 10-item Big-Five questionnaire as well as a questionnaire on integrity. Table 9 shows marginal effects of personality traits measures and integrity for the treatments “No audit” and “Audit-and-fine” separately and for both combined. Effects of risk aversion are analyzed for the Audit-and-fine treatment only, as risk attitudes have only been elicited in this treatment. Note that all characteristics in Table 9 are scaled from -1 to $+1$. The coefficients thus reflect the effect of a change from the theoretical minimum to the (neutral) midpoint or from the midpoint to the theoretical maximum.

For personality traits, we do not find systematic effects. For example, individuals who are

Table 9: Effect of personality traits, integrity, and risk attitudes on upcoding

	Dependent variable: upcoding			
	(1) No Audit	(2) Audit-and-fine	(3) Both	(4) Audit-and-fine
Agreeableness	-0.072 (0.083)	-0.059 (0.124)	0.043 (0.099)	-0.009 (0.115)
Conscientiousness	0.477** (0.170)	-0.048 (0.118)	-0.013 (0.089)	-0.016 (0.116)
Extraversion	0.082 (0.119)	-0.106 (0.078)	-0.112 (0.078)	-0.060 (0.074)
Neuroticism	-0.542** (0.258)	-0.069 (0.090)	-0.028 (0.091)	-0.051 (0.087)
Openness	-0.100 (0.108)	-0.090 (0.095)	-0.183** (0.067)	-0.112 (0.096)
Integrity	-0.163 (0.124)	-0.241 (0.149)	-0.312** (0.125)	-0.201 (0.156)
Risk aversion				-0.058 (0.075)
N decisions	220	270	490	270
N subjects	44	54	98	54

Notes: This table shows average marginal effects based on logit models. Standard errors clustered on subject level in parentheses. We consider decisions only where upcoding is possible, which means we exclude entries for $w_j = 1$, 200g. In the regressions, we control for gender, medical major, four age categories (18-20, 21-25, 25-29, 30+), and for the true birth weights shown. In Model (4), we also control for social trust. All personality characteristics are measured on a scale from -1 to $+1$. The coefficients thus reflect the effect of a one-unit change; for example, a change from the theoretical minimum to the (neutral) midpoint or from the midpoint to the theoretical maximum. * $p < 0.1$; ** $p < 0.05$.

more conscientious, are more likely to upcode birth weights in the No-Audit treatment, but not

in the Audit-and-fine treatment. More neurotic individuals are *less* likely to upcode without audit, but not with audit; see Models (1) and (2) of Table 9. More open individuals are less likely to upcode, with a significant coefficient only for the combined data set (Model 3).

As to integrity, we find that individuals with higher integrity are less likely to misreport birth weights. The estimates are negative and of similar orders of magnitude for all model specifications in Table 9 and significant at a 5 percent level for the combined data set. In particular, individuals with higher integrity are 32 percentage points less likely to upcode, see Model (3) of Table 9. This result has a meaningful interpretation and we thus further investigate the effect of audits on upcoding when differentiating between individuals with different levels of integrity. Table 10 shows estimation results for individuals split at the median level of integrity. We find that individuals with low integrity are somewhat more likely to upcode in the No-Audit treatment (92 versus 82 percent). Audit-and-fine reduces the likelihood of upcoding for individuals with low and high integrity in a very similar magnitude of about 21 percentage points. The same holds true when differentiating between choices where detection and no detection of misreporting is possible under audit.

For risk aversion, coefficients indicate that when individuals are more risk-averse they are slightly less likely (6 percentage points) to misreport birth weights. While intuitively meaningful, this finding is not significant, however; see Model (4) of Table 9.

Table 10: Effect of Audit-and-fine on the probability of upcoding by integrity

	Dependent variable: upcoding			
	(1)	(2)	(3)	(4)
	Integrity <Median	Integrity >Median	Integrity <Median	Integrity >Median
Mean dep. var (no audit)	0.922	0.819	0.902	0.798
Audit-and-fine	-0.214** (0.066)	-0.208** (0.084)		
Audit-and-fine \times detection			-0.312** (0.095)	-0.308** (0.102)
Audit-and-fine \times no detection			0.076 (0.092)	0.132 (0.096)
N decisions	245	245	196	196
N subjects	49	49	49	49

Notes: This table shows average marginal effects based on logit models. Standard errors clustered on subject level in parentheses. We consider decisions only where upcoding is possible, which means we exclude entries for $w_j = 1, 200g$. Individual control variables: four age categories (18-20, 21-25, 25-29, 30+), sex, field of study (medicine vs economics), Big Five inventory. In Models (3) and (4), the data for 1,500g are excluded due to the detection possibility being endogenous. * $p < 0.1$; ** $p < 0.05$.

Result 5. *We find no evidence that (a) personality traits and (b) risk attitudes affect individuals' upcoding in a robust and systematic way. (c) We find some evidence that the likelihood of upcoding under No-Audit and Audit-and-fine decreases with the integrity of individuals.*

6 Discussion

In this section, we discuss our findings and further relate them to the literature. First, we find that the vast majority of the subjects' decisions in our baseline treatment reflect dishonesty (Result 1). Introducing a random audit (10% detection probability) coupled with a fine significantly reduces the likelihood of upcoding (Result 2). One might argue that *audits* as such without fining will be sufficient for inducing honest behavior. Theoretical argumentation and empirical evidence suggest that this seems not to be the case. From a theoretical perspective, a fine is assumed a necessary incentive to induce compliance and is, thus, included in theoretical audit models (e.g., Becker 1968, Clark et al. 2004, Kuhn and Siciliani 2013). From an empirical viewpoint, Lindeboom et al. (2016) study the effect of changes in audit rates without a fine on Dutch health care providers' behavior and report almost no evidence for behavioral responses.

Findings in Germany also run counter to this argument: audits in German hospitals are conducted by the National Association of Statutory Health Insurance Funds (GKV) in cooperation with the regional medical review boards.¹¹ GKV reported in 2013 that at that time hospitals did not bear the risk of any sanctions if they were caught making incorrect claims (GKV-Spitzenverband, 2013). They only had to return the amount incorrectly claimed. Moreover, if the audit did not result in any reduction of the reimbursement claim, GKV even had to pay an expense allowance to the hospital for preparing the 'unsuccessful' auditing. As incorrect claims rose from 35% in 2006 to more than 50% in 2011,¹² the regulations in Germany did not seem to provide appropriate incentives for hospitals to change their practices. In contrast, as GKV pointed out, they protected those hospitals that submitted incorrect claims. GKV heavily complained and insisted that effective sanctions had to be established by law to reduce the number of incorrect billings and thus the loss to the insurers and the insured. Since then, law regulations did change (GKV-Spitzenverband, 2017), even though they are not as far-reaching as, for instance, in the US.

Second, a further concern might relate to using *student participants* in our experiment. Even though recruiting practicing physicians is a promising option planned for future research, student participants provide valid information as well. On the one hand, medical students are future physicians and economics students are future managers probably also in hospitals. Moreover, it has been shown that dishonest student behavior is a predictor for behavior in later work envi-

¹¹They send teams to randomly selected hospitals to evaluate the coding and treatment of patients by auditing patients' medical records.

¹²As a result of incorrect claims, the statutory health insurance funds lose about 2 billion Euro per year (GKV-Spitzenverband, 2013).

ronments. For instance, Sims (1993) reports that subjects who admitted to having engaged in academic misconduct also admitted to a wide range of unethical business practices. In addition, those subjects who engaged in behavior considered severely dishonest in college also engaged in behavior considered severely dishonest at work. Sierles et al. (1980) found a strong correlation between students cheating in medical school and in patient care. Finally, Abeler et al. (2016) report about several recent studies showing that behavior in lying experiments significantly correlates with cheating behavior outside the lab; see, for example, Cohn and Marechal (forthcoming), Dai et al. (forthcoming), Hanna and Wang (2017), and Potters and Stoop (2016).

Third, our finding that *medical students* are rather dishonest under No-Audit might be surprising at first sight. It is, however, in line with self-report studies showing rather high percentages of cheating behavior admitted by medical students (e.g., Sierles et al., 1980, Rennie and Crosby, 2001, Kukolja Taradi et al., 2010). As to the missing impact of gender on dishonest behavior under no audit, our results correspond to those of self-report studies; see, for example, Rennie and Rudland (2003).

Fourth, we find that personality traits are not systematically related to dishonest behavior. What is worth noting, however, is that conscientiousness is significant for increasing and (neuroticism for reducing) upcoding under No-Audit. This is in contrast to Williams et al. (2010) and Giluk and Postlethwaite (2015), who find self-reported academic dishonesty to be significantly negatively correlated with conscientiousness, while the relevance of neuroticism was not supported. This behavioral difference may be explained by the fact that conscientious people are characterized as being industrious, systematic, dutiful, high on achievement striving, and hardworking (Trautwein et al., 2009). In our experiment, participants with these traits can be thought of as those people who strive for high payoffs by upcoding and take into account that their dishonest behavior cannot be detected. In the aforementioned studies, however, conscientious people are not in need of cheating because they can realize their goals by their own skills and organization.

Finally, we relate our study to the body of experimental literature on dishonest behavior. The mere extent of (full and partial) dishonesty in our experiment exceeds reported dishonest behavior, for example in Fischbacher and Föllmi-Heusi (2013), Abeler et al. (2014), and Mazar et al. (2008). Fischbacher and Föllmi-Heusi (2013) find that 39% of subjects are honest and about 20% lie to the fullest extent. In a survey of 72 experimental studies from 43 countries, Abeler et al. (2016) report that subjects forgo on average about three quarters of the potential gains from lying. On the other hand, Abeler et al. (2014), in a representative sample of the German population, find that aggregate reporting behavior is close to the expected truthful distribution. There are several possible explanations for the different degrees of dishonesty found in our paper: (i) The lying experiments above are typically piggybacked to other unrelated experiments. That means subjects earn additional money by taking part in the unrelated experiments, whereas in our experiment the reporting task is the only source of payoff (besides a payment for completing

the post-experimental questionnaire). (ii) In nearly all lying experiments, the researcher does not know whether an individual is lying or not. Only a few studies use a design similar to ours in which the researcher can observe whether individuals are dishonest. When using a double-blind payment procedure as we do, Abeler et al. (2016) find reported numbers to be considerably lower, with only 17% participants lying. In a non double-blind procedure, the authors state the share of subjects misreporting their draw to be lower, but not significantly. Gneezy et al. (2018) find that in the observed treatment the reported numbers are significantly higher than the actual outcomes, with 26% of participants lying.¹³ Notably, the share of dishonest behavior in our experiment is even more pronounced when contrasted with studies using a more germane payment procedure. (iii) Our study uses a neonatology frame and this frame seemed to have served as a justification for dishonest behavior. For instance, 20% of the participants in the No-Audit treatment stated (in an open question) as motives for their decisions that they did not see any problems in upcoding because the early-born baby receives optimal medical care, irrespectively of the DRG paid. About 10% of the participants in the same treatment used arguments regarding the hospital — for example, upcoding would be justified because financial burdens of the hospital are reduced and it thus allows the hiring of additional personnel or the buying of equipment, which in turn benefits the patient in the end. Such motives are absent in the studies on dishonesty mentioned above.

7 Conclusion

We introduce a controlled incentivized behavioral experiment framed in a neonatal care context to analyze the effect of audits coupled with fines on individuals' honesty in a simple birth weight reporting task. The behavioral data elicited using a novel experimental design provide new evidence on upcoding at an individual level. This is particularly due to the distinct feature of our experiment, which allows for a comparison of upcoding with and without the detection probability within the same treatment. Under Audit-and-fine, participants can upcode at some of the birth weights without bearing the risk of being discovered: underreporting of birth weights by small manipulations cannot be detected, while larger manipulations within and between birth weight intervals can.

We find high percentages of upcoding in the baseline No-Audit treatment where detection is not possible. About three quarters of all decisions comprise upcoding. For the true birth weight of 1,500g, nearly 70% of these decisions are fully dishonest, about a quarter are partially dishonest, and only 2% are honest. Introducing Audit coupled with a fine substantially affects the subjects' reporting behavior. Honest birth-weight reports rise significantly, while upcoding is significantly reduced. This is particularly true for detectable upcoding, while no such effect is

¹³In contrast to our experiment, the detection probability seems not to affect dishonest behavior (see, e.g., Mazar et al. 2008, Shu et al. 2012).

found when discovery is not possible. At the 1,500g threshold, participants prefer the safe, but lower, profit to the risky, but maximal return by shifting entries from detectable fully dishonest to non-detectable partially dishonest weights, i.e., detectable fully dishonest entries decrease while honest and partially honest choices rise.

Analyzing the influence of the subjects' characteristics, we find behavioral differences in upcoding behavior with regard to major and gender. Medical students under No-Audit behave slightly more dishonestly than economics students. They are affected more strongly under Audit-and-fine by reducing upcoding substantially. This holds also when detection is not possible. When upcoding cannot be detected, economics students increase upcoding in contrast to medical students, a behavior which implies compensating for forgone opportunities to cheat when detection is possible. At the 1,500g threshold, Audit-and-fine significantly reduces full dishonesty with medical students increasing honest and partially dishonest entries; none of the economics students, however, behave honestly. As to gender, we observe similar upcoding behavior between males and females under No-Audit. Audit-and-fine causes both males and females to reduce upcoding significantly, the decrease being higher for female compared to male students when detection is possible.

Concerning personality traits measured by the Big Five Inventory, we do not find robust effects. While under No-Audit Conscientiousness is significant for increasing and Neuroticism for reducing upcoding, these effects vanish under Audit-and-fine. Analyzing the effect of Integrity, participants scoring high on integrity upcode somewhat less than those who score low in the Baseline treatment. When upcoding can(not) be detected, both reduce (increase) upcoding (in)significantly, but the effect is not different.

The evidence of the existence of dishonesty in healthcare provided by our experimental methodology and the factors that may motivate such behavior are important. The non-verifiability of birth weight manipulation in actual neonatal care settings calls for a research method that allows us to study the behavior, characteristics, and underlying motivations of decision-makers *directly* and at the *individual* level. Currently, false reporting of birth weights can only be revealed indirectly by analyzing aggregate data. Our finding that audit coupled with a fine leads to more honest behavior, and to reduced upcoding when detectable, but that it causes increased dishonesty when detection is not possible highlights two issues. First, audits and fines are important weapons to cope with upcoding. Second, the non-detectability of upcoding birth weights in neonatology inherent in the present reporting system calls for basic changes in the method of recording the birth weight and/or in the DRG reimbursement system to make the system fraud-proof.

Taken together, our results draw attention to the important challenge of understanding how audits and fines affect dishonesty in health care and how dishonesty relates to individuals' characteristics. Finally, an appealing feature of our parsimonious design is that it lends itself to further research on dishonesty in health care. For example, an interesting follow-up question is

how different levels of detection probabilities or fines affect the dishonesty of individuals.

References

- ABELER, J., A. BECKER, AND A. FALK (2014): “Representative evidence on lying costs,” *Journal of Public Economics*, 113, 96 – 104.
- ABELER, J., D. NOSENZO, AND C. RAYMOND (2016): “Preferences for Truth-Telling,” IZA Discussion Papers 10188, Institute for the Study of Labor (IZA).
- BARROS, P. AND G. BRAUN (2017): “Upcoding in a National Health Service: the evidence from Portugal,” *Health Economics*, 26, 600–618.
- BECKER, D., D. KESSLER, AND M. MCCLELLAN (2005): “Detecting Medicare abuse,” *Journal of Health Economics*, 24, 189 – 210.
- BECKER, G. S. (1968): “Crime and Punishment: An Economic Approach,” *Journal of Political Economy*, 76, 169–217.
- BEN-NER, A., F. KONG, AND L. PUTTERMAN (2004): “Share and share alike? Gender-pairing, personality, and cognitive ability as determinants of giving,” *Journal of Economic Psychology*, 25, 581–589.
- BOCK, O., I. BAETGE, AND A. NICKLISCH (2014): “hroot: Hamburg Registration and Organization Online Tool,” *European Economic Review*, 71, 117 – 120.
- BUNDESRECHNUNGSHOF (2011): *Bemerkungen 2010 zur Haushalts- und Wirtschaftsführung des Bundes: Weitere Prüfungsergebnisse. (2010 Annual Report on Federal Financial Management: Further Audit Findings)*.
- CANTRELL, G. (2015): *Fraud in Medicare—Testimony, Deputy Inspector General for Investigations*, US Department of Health and Human Services.
- CARTER, G. M., J. P. NEWHOUSE, AND D. A. RELLES (1990): “How much change in the case mix index is DRG creep?” *Journal of Health Economics*, 9, 411–428.
- CASON, T. N. AND L. GANGADHARAN (2006): “An experimental study of compliance and leverage in auditing and regulatory enforcement,” *Economic Inquiry*, 44, 352–366.
- CHILDS, J. (2012): “Gender differences in lying,” *Economics Letters*, 114, 147 – 149.
- CLARK, J., L. FRIESEN, AND A. MULLER (2004): “The good, the bad, and the regulator: An experimental test of two conditional audit schemes,” *Economic Inquiry*, 42, 69–87.

- COHN, A. AND M. A. MARECHAL (forthcoming): “Laboratory measure of cheating predicts school misconduct,” *The Economic Journal*.
- DAFNY, L. S. (2005): “How Do Hospitals Respond to Price Changes?” *American Economic Review*, 95, 1525–1547.
- DAI, Z., F. GALEOTTI, AND M. C. VILLEVAL (forthcoming): “Cheating in the Lab Predicts Fraud in the Field: An Experiment in Public Transportation,” *Management Science*.
- DOHMEN, T., A. FALK, D. HUFFMAN, U. SUNDE, J. SCHUPP, AND G. G. WAGNER (2011): “Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences,” *Journal of the European Economic Association*, 9, 522–550.
- DONATO, K., G. MILLER, M. MOHANAN, Y. TRUSKINOVSKY, M. VERA-HERNÁNDEZ, ET AL. (2017): “Personality Traits and Performance Contracts: Evidence from a Field Experiment among Maternity Care Providers in India,” *American Economic Review*, 107, 506–510.
- DREBER, A. AND M. JOHANNESSON (2008): “Gender differences in deception,” *Economics Letters*, 99, 197–199.
- DUFWENBERG, M. AND M. A. DUFWENBERG (2018): “Lies in disguise—A theoretical analysis of cheating,” *Journal of Economic Theory*, 175, 248 – 264.
- ELLIS, R. P. AND T. G. MCGUIRE (1993): “Supply-Side and Demand-Side Cost Sharing in Health Care,” *Journal of Economic Perspectives*, 7, 135–151.
- FISCHBACHER, U. (2007): “Z-tree: Zurich Toolbox for Readymade Economic Experiments – Experimenter’s Manual,” *Experimental Economics*, 10, 171–178.
- FISCHBACHER, U. AND F. FÖLLMI-HEUSI (2013): “Lies in disguise—An experimental study on cheating,” *Journal of the European Economic Association*, 11, 525–547.
- FLAHERMAN, V. J., E. W. SCHAEFER, M. W. KUZNIEWICZ, S. X. LI, E. M. WALSH, AND I. M. PAUL (2015): “Early Weight Loss Nomograms for Exclusively Breastfed Newborns,” *Pediatrics*, 135, e16–e23.
- GILUK, T. L. AND B. E. POSTLETHWAITE (2015): “Big Five personality and academic dishonesty: A meta-analytic review,” *Personality and Individual Differences*, 72, 59 – 67.
- GKV-SPITZENVERBAND (2013): *14 Positionen für 2014-Reform der Krankenhausversorgung aus Sicht des GKV-Spitzenverbandes beschlossen vom Verwaltungsrat am 4. September 2013*.
- (2017): *Geschäftsbericht 2016*, Berlin.

- GNEEZY, U., A. KAJACKAITE, AND J. SOBEL (2018): “Lying Aversion and the Size of the Lie,” *American Economic Review*, 108, 419–453.
- GOSLING, S. D., P. J. RENTFROW, AND W. B. SWANN JR (2003): “A very brief measure of the Big-Five personality domains,” *Journal of Research in Personality*, 37, 504–528.
- GREINER, B. (2015): “Subject pool recruitment procedures: organizing experiments with ORSEE,” *Journal of the Economic Science Association*, 1, 114–125.
- GYLFASON, H. F., A. A. ARNARDOTTIR, AND K. KRISTINSSON (2013): “More on gender differences in lying,” *Economics Letters*, 119, 94 – 96.
- HANNA, R. AND S.-Y. WANG (2017): “Dishonesty and Selection into Public Service: Evidence from India,” *American Economic Journal: Economic Policy*, 9, 262–90.
- HARRINGTON, W. (1988): “Enforcement leverage when penalties are restricted,” *Journal of Public Economics*, 37, 29–53.
- HENNIG-SCHMIDT, H. AND D. WIESEN (2014): “Other-regarding behavior and motivation in health care provision: An experiment with medical and non-medical students,” *Social Science & Medicine*, 108, 156 – 165.
- HOUSER, D., S. VETTER, AND J. WINTER (2012): “Fairness and cheating,” *European Economic Review*, 56, 1645 – 1655.
- JANULEVICIUTE, J., J. E. ASKILDSEN, O. KAARBOE, L. SICILIANI, AND M. SUTTON (2016): “How do hospitals respond to price changes? Evidence from Norway,” *Health Economics*, 25, 620–636.
- JÜRGES, H. AND J. KÖBERLEIN (2015): “What explains DRG upcoding in neonatology? The roles of financial incentives and infant health,” *Journal of Health Economics*, 43, 13 – 26.
- KARTIK, N. (2009): “Strategic Communication with Lying Costs,” *The Review of Economic Studies*, 76, 1359–1395.
- KUHN, M. AND L. SICILIANI (2013): “Manipulation and auditing of public sector contracts,” *European Journal of Political Economy*, 32, 251 – 267.
- KUKOLJA TARADI, S., M. TARADI, T. KNEŽEVIĆ, AND Z. OGAŠ (2010): “Students come to medical schools prepared to cheat: a multi-campus investigation,” *Journal of Medical Ethics*, 36, 666–670.
- LINDEBOOM, M., B. VAN DER KLAUW, AND S. VRIEND (2016): “Audit rates and compliance: A field experiment in care provision,” *Journal of Economic Behavior & Organization*, 131, 160 – 173.

- MAZAR, N., O. AMIR, AND D. ARIELY (2008): “The Dishonesty of Honest People: A Theory of Self-Concept Maintenance,” *Journal of Marketing Research*, 45, 633–644.
- MOOKHERJEE, D. AND I. PNG (1989): “Optimal Auditing, Insurance, and Redistribution,” *Quarterly Journal of Economics*, 104, 399–415.
- MUEHLHEUSSER, G., A. ROIDER, AND N. WALLMEIER (2015): “Gender differences in honesty: Groups versus individuals,” *Economics Letters*, 128, 25 – 29.
- POTTERS, J. AND J. STOOP (2016): “Do cheaters in the lab also cheat in the field?” *European Economic Review*, 87, 26–33.
- RAMMSTEDT, B. AND O. P. JOHN (2007): “Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German,” *Journal of Research in Personality*, 41, 203–212.
- REIF, S., S. WICHERT, AND A. WUPPERMANN (2018): “Is it good to be too light? Birth weight thresholds in hospital reimbursement systems,” *Journal of Health Economics*, –.
- RENNIE, S. C. AND J. R. CROSBY (2001): “Are “tomorrow’s doctors” honest? Questionnaire study exploring medical students’ attitudes and reported behaviour on academic misconduct,” *British Medical Journal*, 322, 274–275.
- RENNIE, S. C. AND J. R. RUDLAND (2003): “Differences in medical students’ attitudes to academic misconduct and reported behaviour across the years—a questionnaire study,” *Journal of Medical Ethics*, 29, 97–102.
- SCHLENKER, B. R. (2008): “Integrity and character: Implications of principled and expedient ethical ideologies,” *Journal of Social and Clinical Psychology*, 27, 1078–1125.
- SHIGEOKA, H. AND K. FUSHIMI (2014): “Supplier-induced demand for newborn treatment: Evidence from Japan,” *Journal of Health Economics*, 35, 162–178.
- SHU, L. L., N. MAZAR, F. GINO, D. ARIELY, AND M. H. BAZERMAN (2012): “Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end,” *Proceedings of the National Academy of Sciences*, 109, 15197–15200.
- SIERLES, F., I. HENDRICKX, AND S. CIRCLE (1980): “Cheating in medical school,” *Journal of Medical Education*, 124–125.
- SILVERMAN, E. AND J. SKINNER (2004): “Medicare upcoding and hospital ownership,” *Journal of Health Economics*, 23, 369 – 389.

- SIMS, R. L. (1993): "The Relationship Between Academic Dishonesty and Unethical Business Practices," *Journal of Education for Business*, 68, 207–211.
- TRAUTWEIN, U., O. LUEDTKE, B. W. ROBERTS, I. SCHNYDER, AND A. NIGGLI (2009): "Different forces, same consequence: Conscientiousness and competence beliefs are independent predictors of academic effort and achievement," *Journal of Personality and Social Psychology*, 97, 1115–1128.
- US DEPARTMENT OF HEALTH AND HUMAN SERVICES (2015): *Fiscal year 2015 Agency Financial Report*.
- US DEPARTMENT OF HEALTH AND HUMAN SERVICES AND DEPARTMENT OF JUSTICE (2017): *Health Care Fraud and Abuse Control Program Annual Report for Fiscal Year 2016*.
- WILLIAMS, K., C. NATHANSON, AND D. PAULHUS (2010): "Identifying and Profiling Scholastic Cheaters: Their Personality, Cognitive Ability, and Motivation," *Journal of Experimental Psychology: Applied*, 16, 293–307.

A Appendix

A.1 Instructions

Instructions for the Baseline condition, translated from German. [The text in brackets is for the condition ‘Audit-and-fine’.]

Description of the Experiment

General Information

You are participating in an economic decision-making experiment. Please carefully read through the description of the experiment. It is very important that **you do not talk with other participants for the entire duration of the experiment**. If you violate this rule, you can be excluded from the experiment and will not receive any payment.

If there is anything you do not understand, please take another look at this description of the experiment. If you still have questions, please raise your hand. We will come to you in your cubicle and answer your question personally.

You can earn money in the course of the experiment. The amount of your earnings depends on your decisions. All monetary amounts are displayed in ‘Taler’, at the following rate:

$$1 \text{ Taler} = 1 \text{ Euro Cent.}$$

At the end of the experiment, you will receive the payment you have earned in cash. You will make your decisions on a computer screen in your cubicle. Your decisions cannot be attributed to you personally. All data and answers will thereby be analyzed anonymously.

Decision Situation

The experiment is based on a decision situation in the obstetrics practice of a hospital’s large neonatal care unit. During the experiment, you are entrusted with recording the weight of very early-born infants (birth weights). All participants in this experiment are confronted with the same decision situation.

You have the task of **recording different birth weights, which are displayed on a scale, into a birth report**. The recorded birth weight determines the lump-sum reimbursement per case, which comprises one part of your payment for the experiment.

In total, **six birth weights of different early-born babies will be displayed on the scale, in grams (g)**, on your computer screen. The birth weights displayed can have values between 1,150g and 1,550g. Each birth weight appears only once.

You can record the birth weights in the report in 50-gram intervals, i.e., **1,150, 1,200, 1,250, 1,300, 1,350, 1,400, 1,450, 1,500, or 1,550**.

The birth weight recorded in the report determines which lump-sum reimbursement (in Taler) will be paid for the treatment of an early-born baby. The medical treatment of the early-born babies incurs costs, which should, on average, be compensated by the the lump-sum reimbursement. The costs are dependent on the birth weight displayed on the scale. It is provided for that the early-born babies receive optimal medical care in the neonatal care unit, irrespectively of the lump-sum reimbursement paid.

Lump-sum reimbursements and costs by birth weight are as follows:

Birth weight (in g)	1,150	1,200	1,250	1,300	1,350	1,400	1,450	1,500	1,550
Lump-sum reimbursement (in Taler)	380	380	200	200	200	200	200	120	120
Costs (in Taler)	340	300	260	230	200	180	160	140	130

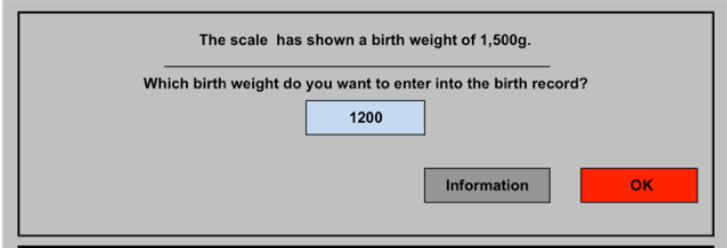
[The newborn will be weighed again the day after birth. This weight will also be recorded in the birth report, by a different person. It should be noted that a newborn baby loses up to 5% of its birth weight in the first 24 hours.

In order to audit hospital bills, the health insurer randomly compares the weights you recorded with the weights measured on the day after the baby’s birth. This audit occurs **after** all birth weights have been recorded **with a probability of 10%**.

If the weight measured on the day after birth is **higher than the recorded birth weight**, the health insurer assumes that the recorded birth weight was too low. This constitutes a serious fake entry.

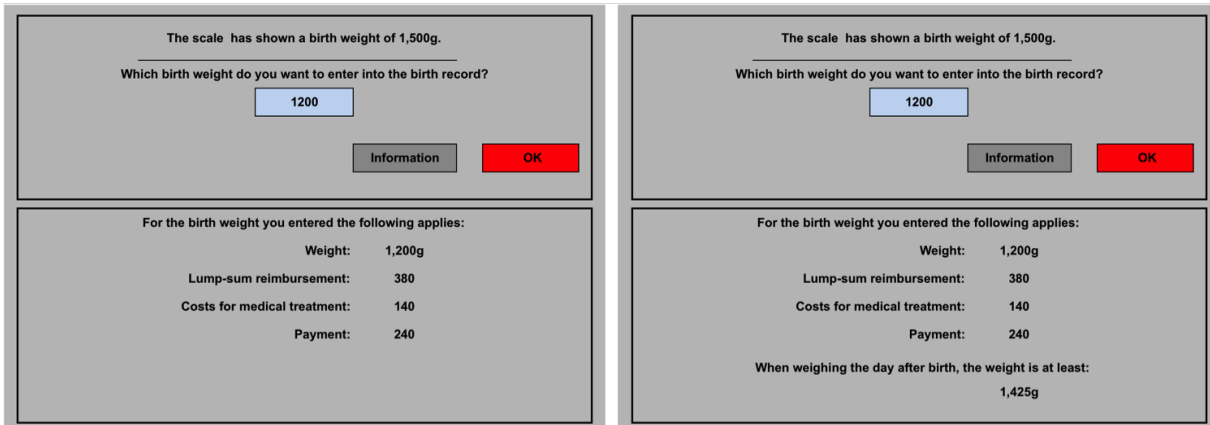
Example: Given a birth weight of 1,500g, the weight 24 hours after birth is at least 1,425g. If the recorded birth weight is, for example, 1,400g, then the health insurer assumes a serious fake entry since the weight on the day after birth is in fact over 1,400g.]

You make your decisions [about what birth weight to record] individually and anonymously on your computer screen. An example of the decision screen is found in the following screen shot:



You enter the birth weight in the designated field. After you have entered the birth weight,

you have the opportunity to learn about the costs arising from the medical care of the early-born baby [and about the minimal weight on the day after birth]. To access this, click on the ‘Information’ button. The following screen shot provides an example:



[The left panel shows the screen for the No-Audit treatment, the right panel for Audit-and-fine.]

You confirm your decision by clicking on ‘OK’ and continue to the next screen.

Payments in the Experiment

Your payment for each recorded birth weight is determined as follows:

Lump-sum reimbursement per case according to the recorded birth weight

minus

Costs for treatment of the early-born baby according to the birth weight displayed on the scale.

After you have made your six decisions, your payment for all entries is calculated as the sum over all six decisions.

Additionally, you will receive a **fixed amount of 400 Taler**. Your total payment is therefore comprised of your payment for the recorded birth weights and the fixed amount.

[In the event that the insurance provider’s audit shows that the weight on the day after birth **for at least one** newborn baby is larger than the recorded weight, this constitutes a serious fake entry. You will therefore be punished with a **reduction in your total payment equaling the amount of your payments for all recorded birth weights, and you will receive the fixed amount only**. That means that the hospital has to treat all children without lump-sum reimbursements. It is not known at the time the treatment is provided whether or not this is the case and therefore has no effect on medical care.

You will learn at the end of the experiment whether the birth weights you entered were audited

and whether a fake entry was discovered.]

Your total payment for the experiment will be converted into Euro and paid out to you at the end of the experiment.

Before the actual experiment begins, we would like to ask you to answer several practice questions that should help to better understand the experiment.

After you have made all of your decisions, we would ask you to answer a series of questions contained in a questionnaire. You receive additional payment for this.

Note:

- You make your decisions anonymously on the computer screen in your cubicle.
- In order that no decision and no payment can be attributed to a certain person, an employee of *BonnEconLab* who is not involved in the experiment will place in your cubicle an envelope with only your cubicle number on it containing the total payment linked to your cubicle.
- **Afterward, please leave the laboratory.**