Appendix A – Additional Information on the Experiment

Appendix A.1 – Sample Screen

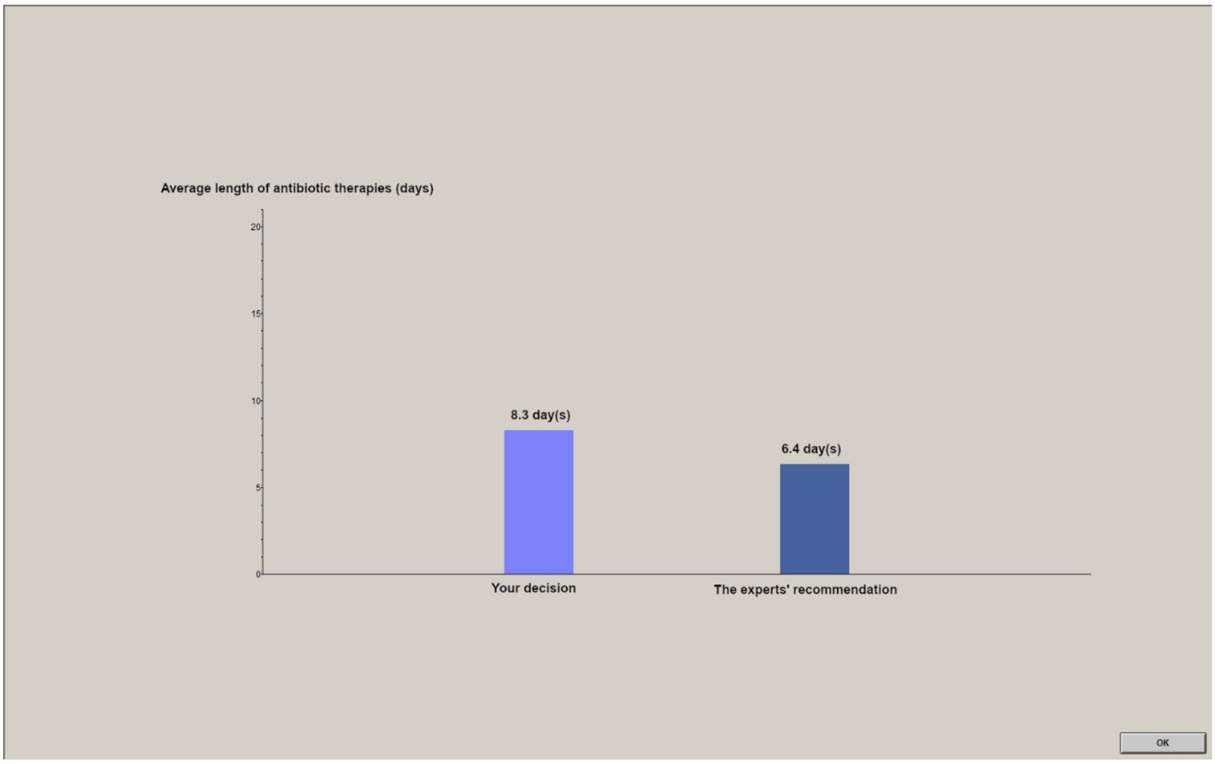

Figure A.1: Sample screen. This figure shows a screenshot of the screen the subjects saw after each round of the experiment (translated from German to English). The left bar shows the average length of therapies the subject chose in the previous round (example), the right bar shows the aggregated expert recommendation. Both average numbers are displayed numerically above the respective bars. The y-axis displays the average length of antibiotic therapies (in days).

Appendix A.2 – Instructions for the Experiment

[Note that in the squared brackets we present instructions from the second and third parts of the experiment.]

You are taking part in a decision experiment. Please read through the instructions carefully. It is important that you do not talk to other participants for the entire duration of the experiment. If you have any questions, please raise your hand. We will come to your cubicle and answer your questions in person.

In this experiment, all monetary amounts are denoted in 'Taler', at a rate of 1 Taler = €1. Your earnings will be paid in cash at the end of the experiment.

You will make your decisions anonymously in your cubicle. All data will be evaluated anonymously. You have drawn your own cubicle number in order to ensure anonymity.

The experiment will last for approximately 60 minutes and consists of three parts. You will receive detailed instructions prior to each stage of the experiment. Please note: Your decisions in each part of the experiment will not have any impact on any other part of the experiment.

At the end of the experiment, you will receive compensation for the experiment.

We also ask you please to answer a few questions at the end of the experiment.

First [Second, Third] part of the experiment

Decision situation

The first part of the experiment relates to a decision situation in the pediatric department of a hospital. You make your decision in the role of the on-duty pediatrician.

In the course of the first part of the experiment, you will be presented with a series of patients, each with different pathologies, symptoms, complaints, or results. If symptoms, complaints, or results are not provided, then they are not considered to be relevant for your decision-making.

In creating an initial treatment plan, you have the task of determining the duration of a course of antibiotics (in days). Here, you can set the length of the course at 0, 1, 2, . . ., 27, or 28 day(s). Note that the respective medicines will be administered according to the relevant guidelines. The initial treatment plan can be adjusted through a reevaluation.

Enter the length of the antibiotics course for each patient in the field 'For how many days do you prescribe antibiotic therapy?' on your computer screen. You can enter whole numbers between zero and 28. Please confirm your decision by clicking 'OK', which will take you to the next screen.

[**After you have made your decision about the length of the antibiotic therapy for all patients, you will be informed about an expert opinion on the average length of antibiotic therapy for patients identical to those for whom you have made treatment decisions. The expert opinion is based on responses from 20 leading pediatricians drawn from a representative sample of children's hospitals in Germany**[*]]

Earnings

For carrying out the task in the first part of the experiment – determining the length of antibiotic therapy for a series of patients – you will receive a fixed payment of 50 Talers.

Important information:

- Make your decisions anonymously on your computer screen.

- In order that no decision or payout can be matched with a particular participant, an employee of the Department of Business Administration and Personnel Economics at the University of Cologne, who is not involved in conducting the experiment, will place in your cubicle an envelope that is marked only with the cubicle number and contains the total payout for your cubicle.

- Afterwards, please leave the room in which the experiment was conducted.

[*] This survey was conducted in August and September 2014 among head physicians in German children's hospitals. Out of a total of 50 randomly chosen German children's hospitals, 20 hospitals answered questions about the length of antibiotic therapy in full. The study is archived in the German Clinical Trials Register under the study number DRKS00006782

Appendix A.3 – The Medical Cases

A.3.1 List of the Medical Cases

Table A.1: Medical cases (with categories of pediatric infectious diseases)

	Case description	Randomized order
Neonatal infections		
1	Newborn at 38 weeks of gestation at the age of four hours after a normal standardized pediatric examination. Spontaneous vaginal delivery, rupture of membranes at birth, maternal fever 38.5 C at birth, C-reactive protein (CRP) < 5 mg/dl (mother), Group B-Streptococcal (GBS) status is negative. The child's vital signs and clinical examination are normal.	4
2	Term newborn at the age of six hours after normal postnatal examination. Spontaneous vaginal delivery, rupture of membranes < 18 hours before the onset of labor, positive maternal GBS status two weeks before birth. No antenatal antibiotic treatment. The child's vital signs and clinical examination are normal.	39
3	Newborn at 40 weeks of gestation at the age of 12 hours after a normal postnatal physical examination. Spontaneous vaginal delivery, rupture of membranes > 18 hours before the onset of labor, positive GBS status two weeks before birth. Maternal antibiotic treatment three hours before birth. The child's vital signs and clinical examination are normal. In the blood test, maximal CRP (C-reactive protein) 18 mg/l and Il-6 (Interleukin 6) 10 ng/l.	20
4	Term newborn on the second day of life. Spontaneous vaginal delivery, rupture of membranes at birth, normal postnatal physical examination. In the clinical examination, the child was hypotonic with gray skin color, impaired microcirculation, tachypnea, and dyspnea. In the blood tests initiated by you, a CRP shows a maximum of 35 mg/l, Il-6 > 8 ng/l. The blood cultures and newborn smears, received after two days, were without pathogen detection.	23
5	Newborn of the 38th gestational week, at the age of two days. Admission to the NICU and start of an antibiotic therapy after an abnormal physical examination. In blood test, maximal CRP 15 mg/l, Il-6 < 8 ng/l. The CSF findings were normal. In the blood culture, detection of Staphylococcus epidermidis. The child's vital signs and physical examination are currently normal.	19
6	Newborn with a gestational age of 39 weeks at the age of 20 hours. In the physical examination, the child is hypotonic with impaired microcirculation and hypothermia. In blood test, CRP > 75 mg/l, Il-6 150 ng/l. The CSF findings are negative. In the blood culture detection of Staphylococcus epidermidis.	24
7	Newborn at 41 weeks of gestation, at the age of five days. In the clinical examination, the infant shows hyperexcitability and a gray skin color, tachypnea, dyspnea and fever (max. 39°C). In the laboratory analyses initiated by you, the CRP is 90 mg/dl, and the interleukin 6 (Il-6) is 1,450 ng/l. In cerebrospinal fluid (CSF), there were 80 leukocytes/µl. The culture of the CSF remained negative. In the blood culture, E. coli was detected.	5
8	Preterm infant with spontaneous vaginal delivery after 32 weeks of pregnancy. Prenatal maternal antibiotic prophylaxis and a history of rupture in the 29th week of gestation. Mother GBS status negative. Initially slight respiratory distress syndrome. The patient is stabilized by nasal continuous positive airway pressure (nCPAP) quickly, CRP < 5 mg/dl, Il-6 < 8 ng/l. The respiratory support could be terminated at the second day of life.	13
9	Preterm infant after spontaneous vaginal delivery in the 33rd week of pregnancy. Rupture of membranes at birth, positive maternal GBS status, and antenatal IV antibiotic treatment three hours before birth. Initial slight respiratory distress syndrome. The patient rapidly stabilizes under nCPAP. The ventilatory support can be terminated at the second day of life. Initiation of the antibiotic therapy in the delivery room. In the blood test, initiated by you on the second day of life, CRP 15 mg/l and Il-6 < 8 ng/l.	31
10	Twin preterm infant at the 32nd week of gestation. Spontaneous vaginal birth. The GBS-positive mother received an intravenous antibiotic treatment six hours before birth. Initial respiratory distress syndrome (III°). Surfactant application and further respiratory support with nCPAP in the first hours of life. Initiation of an antibiotic treatment in the labor ward. CRP 30mg/l, Il-6 120 ng/l. Blood cultures and neonatal smears were negative.	15

11	Premature infant with a gestational age of 28 weeks. Caesarean section due to maternal HELLP syndrome. Initial slight respiratory distress syndrome. Rapid stabilization of the respiratory state under nCPAP. Implantation of a silastic catheter. On the fifth day of life deterioration of general condition, gray patchy skin color, capillary refill prolonged and increasing oxygen demand. Removal of the catheter. Improvement of the clinical condition after application of an antibiotic therapy. In blood, maximal CRP 35 mg/l and Il-6 148 ng/l. The blood cultures and newborn smears were negative.	32
12	Premature infant at the 25th week of gestation after three cycles of antibiotic therapy because of systemic in inflammatory response syndrome (SIRS) and catheter sepsis. O2-supply via nasal prongs, oral nutrition. At the age of eight weeks, poor feeding, vomiting, and abdominal distension. With suspected septicemia or necrotizing enterocolitis initiation of an antibiotic treatment. The CRP value was 35 mg/l, Il-6 was 480 ng/l. The blood cultures were negative. In the neonatal smears, detection of Staphylococcus epidermidis, Enterobacter species, and Candida albicans. Immediate improvement of the clinical condition after the initiation of the therapy.	17
Infections of the CNS		
13	Six-year-old boy with sudden fever between 39°C and 40°C. His temperature cannot be reduced with physical and pharmacological measures. Severe headaches, neck pain, and vomiting. Admission with suspected meningitis and implementation of an antibiotic treatment. In CSF: turbid appearance, leukocyte count > 1,000/µl. CSF culture: negative.	27
14	Eight-year-old girl with severe headache and neck pain. High fever up to 40°C since the previous day. By suspected meningitis admission in your clinic and initiation of an antibiotic therapy. CSF results: turbid, cell count > 1,000/µl. In the rapid test and in the CSF culture, detection of meningococcus.	7
15	Ten-year-old boy with infection of the respiratory tract for one week. Fever up to 39°C, headache, and photophobia since the previous night. Admission to the hospital with suspected meningitis. CSF findings: cell count > 1,000/µl, protein 500 mg/l, lactate 4.5 mmol/l. Pneumococcus species were detected in the blood culture.	6
16	Two-year-old former premature infant with ventriculoperitoneal shunt. High fever up to 40°C, drowsiness, and vomiting since the previous day. CSF after puncture of the shunt valve: cell count > 1,000/µl. In CSF, detection of Staphylococcus. The ventriculoperitoneal shunt was explanted shortly after admission.	38
Bone and joint infections		
17	12-year-old boy with pain in his left foot since the previous day. Pain when standing, redness and swelling and effusion in the area of the ankle. Trauma history negative and no visible external injury. Hospital admission for puncture and antibiotic therapy. In the puncture, detection of Staphylococcus aureus. Significant improvement of the clinical symptoms and normalization of the inflammation parameters within the first week of antibiotic treatment.	10
Upper respiratory tract infections		
18	Three-year-old child with acute ear pain, infection of the upper respiratory tract, serous rhinitis, and a maximal body temperature of 38.5°C. Otoscopy: redness and withdrawal of the tympanic membrane.	1
19	Eight-month-old infant in poor general condition. Apparent ear pain until the day before. Infection of the upper respiratory tract with purulent rhinitis and temperature up to max. 40°C. Otoscopic findings: purulent otorrhea with perforated eardrum.	12
20	Seven-year-old child with ear pain, infection of the upper respiratory tract, serous rhinitis, and fever up to max. 40°C for three days. Otoscopic findings: redness of the eardrum.	26
21	Ten-year-old girl in good general condition with serous rhinitis and coughing for one week. Frontal headache when tilting the head since the previous day.	3
22	12-year-old girl in good general condition with serous rhinitis and cough for two weeks. Severe facial pain for five days. Fever > 39°C during the clinical examination.	16
23	Eight-year-old boy with purulent rhinitis and cough for one week. Fever > 39°C and strong frontal headache for two days.	35
24	Eight-year-old boy with fever up to 39.8°C, fine maculate, slightly elevated, pale red rash, glossitis, and erythematous tonsils. Positive streptococcal rapid test.	36
25	Five-year-old girl with difficulty in swallowing, red tonsils, and swelling of the cervical lymph nodes without fever. Positive streptococcus A rapid test.	2

26	Ten-year-old girl with rapidly rising fever, pain and malaise. The tonsils are swollen and red, and there is a cervical lymph node swelling. Streptococcal A rapid test positive.	18
Urinary tract infections		
27	Detection of bacterial species > 10^5/ml in the investigation of the midstream urine of a 13-year-old female adolescent. The routine clinical examination was unremarkable.	30
28	15-year-old girl with dysuria, pollakiuria, and temperature up to 38.5°C. In the urinary analysis, leukocyturia and bacteriuria.	9
29	16-year-old girl with frequent, imperative urinary urgency and hematuria for two days. On the day of examination, strong malaise, fever up to 40.5°C and flank pain. In the urine analysis, 3,000 leukocytes/µl, massive bacteriuria, and 300 isomorphic erythrocytes/µl. In the blood, 19,000 leukocytes/µl and CRP 120 mg/l. In the ultrasound examination, the kidneys were normal and there was no urinary obstruction.	21
30	15-year-old girl with dysuria for the first time, pollakiuria, flank pain, and fever up to 40°C. On the urine strip test (midstream urine) leukocytes ++, nitrite ++. In blood test leukocytosis and a CRP value of 100 mg/l. In the ultrasound examination, the left kidney was enlarged and partly echogenic, no urinary obstruction.	11
31	Four-month-old male infant with fatigue and fever up to 40.5°C. The CSF findings were normal. In the urine probe after catheterization: 500 leukocytes/µl. In blood test, leukocytes 24,000/µl and CRP 80 mg /l. The renal ultrasound examination revealed a suspected reflux.	22
32	Five-month-old male infant with fever up to 40°C. Poor general condition without a clear infectious focus. In the blood test: 16,400 leukocytes/µl , CRP 95 mg/l. Urine test strip after bladder puncture: leukocytes +++, erythrocytes ++, nitrite +, proteins +. Urine culture: Detection of E. coli 10^6/ml.	8
Lower respiratory tract infections		
33	A six-week-old infant has been suffering from rhinitis for three days, fever up to 38°C, and increasingly dry cough. The child is pale, with nasal flaring, tachypnea, dyspnea, and subcostal chest retractions. Bilateral attenuated respiratory sound, fine crackles, and expiratory wheezing. The rapid test for respiratory syncytial virus (RSV) is positive. In the blood test, 5,300 leukocytes/µl and CRP 20 mg/l.	34
34	A nine-month-old female infant has been suffering from fever up to 38°C for one week, rhinitis, and dry cough. Symptomatic therapy with suspected viral infection. Since the previous day, deterioration of the general condition and fever up to 40°C. Bilateral attenuated breath sounds and occasional fine crackles and expiratory wheezing in the auscultation. The RSV rapid test is positive. In blood, leukocytes 15,000/µl, CRP 70 mg/l.	40
35	Three-month-old infant with tachypnea and cough resembling whooping cough. Postnatal purulent conjunctivitis. Chlamydia trachomatis pneumonia is suspected.	14
36	Five-year-old boy with fever up to 39.5°C and abdominal pain. Auscultation: inspiratory fine crackles and attenuated breath sounds. Laboratory findings: leukocyte 27,800/µl, CRP 38 mg/dl. The x-ray reveals a lobar pneumonia.	28
37	Seven-year-old girl with severe abdominal pain and fever up to 39°C. In the physical examination: basal attenuated breath sounds in the auscultation and basal damping in the percussion; the abdomen is normal. In blood, 13,500 leukocytes/µl and CRP 77 mg/l. The chest x-ray revealed pneumonia.	37
38	Six-year-old girl with severe cough, purulent rhinitis, and fever up to 40°C. In the auscultation, fine inspiratory crackles and expiratory wheezing. Laboratory findings: 17,500 leukocytes/µl, CRP 100 mg/l. Bronchopneumonia in the chest x-ray.	25
39	Six-year-old boy with fever up to 40°C and abdominal pain for ten days. The chest x-ray shows pneumonia with basal pleural effusion. After a seven-day-long antibiotic treatment duration, relapse of fever, and occurrence of increasing dyspnea. In the chest x-ray, an abscessing pneumonia is suspected. Surgical application of an abscess drainage.	29
40	Ten-year-old girl with intermittent fever up to 40°C, cough and rhinitis. A therapy with cefuroxime has not lead to an improvement. The chest x-ray reveals central infiltrates with the involvement and compression of the hilum. An atypical pneumonia is suspected.	33

Notes. This table shows the 40 medical cases used in the expert survey and in the experiment. It also shows the six categories of infectious diseases to which the cases can be assigned. The last column provides the randomized order of the cases used in the survey and in the experiment.

A.3.2 Development and Validation of the Cases

The cases had been developed by the clinicians in the research team, (three pediatricians with different sub-specializations) based on textbooks, clinical case reports, and clinical experience (including experience from discussions in regular case conferences).

Afterwards, the cases were validated by five pediatricians of the Department of Pediatrics at the University Hospital Cologne with different sub-specializations (neonatology, infectious diseases, nephrology, neurology, pneumology) and different levels of clinical experience. We asked them to assess the cases with regard to (i) their clarity and comprehensibility, (ii) their relevance in clinical practice, (iii) their plausibility, and (iv) the correctness and completeness of the given information.

As for some infectious diseases, the appropriate length of therapy differs depending on the choice of the antibiotic agent and the dosage; we asked the participants in our study to consider the standard antibiotic agent and the standard dosage for each case when deciding on the length of first-line antibiotic therapy. Therefore, we made sure that each case description comprised all information necessary to determine (an initial clinical diagnosis and) a standard antibiotic agent and dosage. As part of the validation process, we asked the five pediatricians to decide on the length of the therapies and on the agents and dosages they would choose. The case scenarios and all discrepancies in treatment decisions were discussed among the five pediatricians and the research team. For some of the cases, we changed the wording to prevent any misinterpretation of the given information. For some cases, we added further information to rule out any possible differential diagnoses, which were the main reasons for heterogeneous antibiotic treatment decisions made by the five physicians. Furthermore, we matched each case description with the respective treatment recommendation from the handbook published by the German Society for Pediatric Infectious Diseases.[1] By doing so, we made sure that the handbook provided, based on explicitly stated standard antibiotic agents and dosages,a a recommendation on the length of the first-line therapy for each case. This ensured comparability between the decisions from the expert survey and the recommendations from the German Society for Pediatric Infectious Diseases.

[a] For the cases for which several antibiotic agents were recommended, all agents except for the standard agent had to be declared as alternatives to be used only in exceptional cases (e.g., in case of resistance or allergies).

Appendix A.4 – Survey with Directors of German Pediatric Departments

A.4.1 Descriptive Statistics

In total, 20 directors of 50 randomly selected pediatric departments participated in our online survey. The expert sample comprised 19 male pediatricians and one female pediatrician, who were aged between 40 and 62 years. The aggregated expert opinion, i.e., the average length of antibiotic therapy the experts chose for the 40 cases, was 6.42 days (SD 4.94, 95% CI 4.26 to 8.59). This aggregated value served as the 'expert benchmark' in our experiment. See Table C.2 for detailed results of the expert survey.

A.4.2 Comparison with Guidelines

We compared the experts' decisions with published recommendations on the length of antibiotic therapy for each respective case. For this comparison, we only considered recommendations on the length of first-line therapy with the standard antibiotic agent to assess the experts' compliance with recommendations, because the participants in our study were asked to decide on the length of first-line antibiotic treatment with the standard antibiotic agent. We primarily used the recommendations published by the German Society for Pediatric Infectious Diseases.[1] The handbook published by this society provides (based on the use of explicitly stated standard antibiotic agents) a recommendation on the therapy length for each case we used in our study. Moreover, it reflects the consensus of several leading German pediatricians, which leads us to assume that it also reflects local standards of care in pediatric medicine.[b]

Using Fisher-Pitman permutation tests for paired replicates, we analyzed whether the decisions made in the expert survey were significantly different from the recommendations. For each case, we compared the 20 decisions of the experts with the range of recommended numbers of treatment days. We considered decisions as compliant with the recommendations if they were within the range of recommended numbers of treatment days or deviated one day at most (i.e., the recommended intervals were extended by +/-one day). In doing so, we adopted the measure of compliance with recommendations on the length of antibiotic therapy that has been applied by other scholars.[2] The interval was not extended by one day, however, if no antibiotic therapy (zero days) or an explicit maximum or minimum number of days is recommended (e.g., for the recommendation 'from one day up to a maximum of two days', we

[b] Note that the recommendations are very similar to recommendations from national and international guidelines. The handbook is an aggregate of available evidence, which should also be included in those guidelines.

accepted one or two days as compliant with the recommendation. For ‘at least 10 days up to 14 days’, the range between 10 and 15 days was considered appropriate. Note that we did not extend the interval to zero days if the lower boundary is one day). For cases for which the handbook provides no upper or lower boundary (e.g., ‘at least 10 days’), we used recommendations from further national and international guidelines as references (see Table A.2 for details).

In 80 percent of the cases, the experts’ decisions were in line with the recommendations, i.e., only in eight out of the 40 cases (20%) did the decisions significantly differ from what guidelines recommend (with a p-value < 0.05). Comparable studies reporting compliance rates with antibiotic prescribing guidelines are rare. Labenne et al.,[2] which is, to the best of our knowledge, the only study that examines guideline compliance regarding the length of antibiotic therapy for children, reported a compliance rate of 70 percent. Other studies, which lack comparability since they do not consider length of antibiotic therapy, found low average medical guideline compliance rates among physicians of 61 percent[3] or 54.5 percent.[4] Given the experts’ large guideline compliance rate in our survey, we argue that the aggregated expert opinion can be considered a suitable benchmark for an appropriate length of antibiotic therapy.

Table A.2: Results of the expert survey (n=20) and recommendations on length of therapies

Cases (ordered by category)	The experts' decisions in days (n=20)					Recom-mended length of therapy (in days)	Absolute deviation of the experts (n=20) from the recommendations (in days)				
	Min	Max	Mean	Median (IQR)	s.d.		Min	Max	Mean	Median (IQR)	s.d.
Neonatal infections											
1	0	5	0.65	0 (0-0)	1.63	0	0	5	0.65	0 (0-0)	1.63
2	0	3	0.15	0 (0-0)	0.67	0	0	3	0.15	0 (0-0)	0.67
3	0	10	2.40	0 (0-5)	3.12	1-2 (max.)	1	8	2.05	1 (1-3)	1.93
4	0	10	5.50	5 (5-7)	2.42	5 (-7)	0	4	0.60	0 (0-1)	1.10
5	0	7	2.80	3 (0-5)	2.46	1-2 (max.)	0	5	1.85	1 (1-3)	1.42
6	3	10	6.85	7 (5-9.25)	2.16	7 (-10)	0	3	0.50	0 (0-1)	0.76
7	7	21	11.65	10 (7.5-14)	4.69	21	0	13	8.50	10 (6-12.5)	4.38
8	0	3	1.20	0 (0-3)	1.40	0	0	3	1.20	0 (0-3)	1.40
9	0	10	3.80	3 (2-5)	3.00	1-2 (max.)	0	8	2.40	1 (1-3)	2.39
10	2	10	5.25	5 (3.5-6.5)	2.12	5 (-7)	0	2	0.50	0 (0-1)	0.76
11	2	10	6.30	6 (5-7)	2.03	5 (-7)	1	6	2.50	2.5 (1-3)	1.76
12	5	14	7.35	7 (5.5-7.75)	2.23	5 (-7)	0	6	0.60	0 (0-0)	1.47
Infections of the CNS											
13	3	21	9.70	10 (7-10)	3.66	7-10	0	10	0.95	0 (0-0)	2.39
14	5	21	8.35	7 (7-9.5)	3.92	4-7	0	13	1.45	0 (0-1.5)	3.30
15	7	21	10.50	10 (7.25-13)	3.47	7-10	0	10	1.10	0 (0-2.25)	2.43
16	5	21	11.85	14 (10-14)	3.73	at least 10-14	0	6	1.00	0 (0-2.25)	1.89
Bone and joint infections											
17	7	28	17.45	17.5 (14-21)	6.58	21	0	13	5.00	6 (0-6)	4.36
Upper respiratory tract infections											
18	0	7	1.60	0 (0-4.5)	2.62	0	0	7	1.60	0 (0-4.5)	2.62
19	5	14	7.20	7 (5-7)	2.28	10	0	4	2.35	2 (2-4)	1.35
20	0	7	1.35	0 (0-3.75)	2.43	5-7	0	4	3.00	4 (1-4)	1.78
21	0	10	1.25	0 (0-0)	2.75	0	0	10	1.25	0 (0-0)	2.75
22	0	14	5.65	6 (5-7)	3.63	10 (-14)	0	9	3.70	3 (2-4)	3.03
23	0	14	5.80	7 (1.25-9.25)	4.09	10 (-14)	0	9	3.65	2 (0.5 -7.75)	3.45
24	3	10	7.85	7 (7-10)	2.23	10	0	6	1.60	2 (0-2)	1.79
25	0	10	5.00	6 (0-7)	3.83	5	0	10	5.00	6 (0-7)	3.83
26	0	10	6.95	7 (5-10)	2.67	10	0	9	2.35	2 (0-4)	2.32
Urinary tract infections											
27	0	5	0.35	0 (0-0)	1.14	0	0	5	0.35	0 (0-0)	1.14
28	0	7	3.50	3 (3-5)	2.01	3 (-5)	0	2	0.40	0 (0-1)	0.68

29	5	14	7.65	7 (7-9.25)	2.16	(7-) 10	0	3	0.30	0 (0-0)	0.73
30	7	14	8.40	7 (7-10)	1.96	(7-) 10	0	3	0.15	0 (0-0)	0.67
31	5	14	8.85	10 (7-10)	2.5	10-14	0	4	1.10	0 (0-2)	1.37
32	7	14	8.60	7 (7-10)	2.3	10 (-14)	0	2	1.20	2 (0-2)	1.01
Lower respiratory tract infections											
33	0	7	0.35	0 (0-0)	1.57	0	0	7	0.35	0 (0-0)	1.57
34	0	10	5.85	7 (5-7)	2.85	7	0	6	1.30	0 (0-1.75)	2.13
35	0	21	12.00	14 (10-14)	4.81	at least 10 (10-14)	0	10	1.50	0 (0-2.25)	2.91
36	0	14	8.00	7 (7-10)	2.88	7	0	6	1.50	1.5 (0-2)	1.79
37	5	14	8.10	7 (7-10)	2.49	7	0	6	1.10	0 (0-2)	1.86
38	5	14	7.55	7 (7-9.25)	2.24	7	0	6	0.90	0 (0-1.75)	1.45
39	5	21	13.45	14 (10-14)	4.77	at least 21 (21-28)	0	16	7.55	7 (7-11)	4.77
40	3	14	9.90	10 (10-10)	2.86	10	0	6	1.30	0 (0-3)	1.81

Notes. This table shows the experts' decisions on the length of antibiotic treatment for each case, as well as the recommendations published by the German Society for Pediatric Infectious Diseases[1] and the experts' absolute deviation from the recommendations. The experts' decisions on the length of therapy, aggregated over all cases, were used as the 'expert benchmark' in our experiment. We assessed the experts' compliance with the recommendations by comparing the experts' decisions with the recommended length of therapy for each case. We allowed a deviation of one day from the recommended number of days (+/- 1 day). We did not allow a deviation if the recommendation is exactly zero days or if an explicit upper or lower boundary is recommended ('at least' or 'max.'). Note that for cases 35 and 39, the recommendations of the German Society for Pediatric Infectious Diseases provide no upper boundary. They recommend *at least 10 days* for case 35 and *at least 21 days* for case 39. To get an upper boundary, we used recommendations from further guidelines. For case 35, we set the upper boundary to 14 days, since the American Academy of Pediatrics recommends 14 days of antibiotic therapy.[5] For case 39, the upper boundary was set to 28 days because the Centers for Disease Control and Prevention (CDC), the American Academy of Pediatrics, and the Pediatric Infectious Diseases Society recommend a length of therapy between 18 and 28 days.[6] To determine the mean absolute deviation of the experts from the recommendations, we analyzed each single decision. If the chosen length of therapy was below the lower boundary of the recommended interval, we calculated the absolute deviation from the lower boundary; if the chosen length of therapy was above the upper interval boundary, we calculated the absolute deviation from the upper boundary. For all decisions that were within the interval of recommended length of therapy or exactly the same as the recommendation (if recommendation is not an interval), the absolute deviation was determined to be zero. For each case, we determined the mean absolute deviation from the recommendations by averaging the absolute deviations across all experts.

Appendix A.5 – Some Photographs from the Experiments

Figure A.2: Some impressions from the experimental sessions. This figure shows the cubicles of the mobile computer laboratory. The left picture shows cubicles of the mobile laboratory at the Department of Pediatrics at the University Hospital Cologne. The middle picture shows parts of the laboratory at the Children's Hospital of the City of Cologne. The right picture indicates the laboratory during the annual conference for pediatricians in Cologne (Päd-Ass 2015).

A.6 Sample Size Calculations and Power Analyses

A.6.1 A-priori Sample Size Calculations

To calculate the required sample size for the detection of a between-subject effect of feedback, we considered the changes in length of antibiotic therapy (measured in days) between Stage 2 and Stage 3 and compared the changes in the intervention group (where feedback was provided) with the changes in the control group. We reviewed the existing literature for a prior to use for our sample size calculation. A recent Cochrane review by Davey et al.[7] summarizes the effect of different interventions to improve antibiotic prescribing practices for hospital inpatients. This review reports a weighted mean reduction of 1.95 days in total duration of antibiotic treatment (95% CI -2.22 to -1.67) associated with the interventions in 14 RCTs. This equals a mean reduction by 28 percentage points and provides a prior for us to determine what change in length of therapy through our feedback mechanism can be considered meaningful. Yet, the average effect is rather large, and so are the effects of most studies included in the review. Moreover, all 14 RCTs were conducted in a hospital setting, which is why the infectious disease cases considered in the review might require longer treatment courses on average than the cases we used in our experiment. Hence, there might be a greater scope for adaption in length of therapies. In this light, we aimed at detecting a change through the provision of feedback which is smaller than 1.95 days.

Instead of comparing decisions on therapy length made in the two experimental groups, we compare changes between the experimental stages that happen in the two groups, because we did not know beforehand how subjects would decide in the first stages. Considering the changes in both groups to measure the effect of our feedback intervention did not require knowing the start values. The effect of providing feedback was defined as the change in the average length of therapies between Stage 2 and Stage 3 in the intervention group compared to the respective change in the intervention group. We consider an average difference of 0.5 days in the change as the minimum relevant effect that should be detected with a sufficient statistical power. This is conservative in light of the large effects found in other studies.[7]

Using Cohen’s d as an effect-size statistic and assuming a standard deviation of 0.65 for the change in both groups, this results in an effect size of 0.769, which we aimed at detecting with a power of 80% (β=0.2) and with an alpha of 0.05. We used G*Power[8] for a two-tailed Mann-Whitney-U test to estimate the required sample size for the detection of a between-subject effect of feedback. In G*Power, the sample size required for a non-parametric test is determined by multiplying the sample size calculated for an equivalent parametric test by a correction factor, referred to as the asymptotic relative efficiency (ARE). We used the ARE method that defines the power of the Mann-Whitney-U test relative to the two groups t-test and chose the most conservative estimation strategy by setting the ARE to its theoretical minimum, although this resulted in larger required sample sizes. This yielded a minimum sample size of 32 required for each group to detect a significant difference between the groups with regard to the change from Stage 2 to Stage 3 with a power of 80%.

A.6.2 Post-hoc Power Calculations

Further, we analyzed the level of statistical power achieved, again using the ‘length of therapy (measured in days)’ as a variable of interest and selecting the ARE method (with the ARE set to its theoretical minimum) of a two-sided Mann-Whitney-U test. Changes from Stage 2 to Stage 3 in the treatment group were compared to changes between the same stages in the control group.

The realized sample size in our experiment was n=73, with n=39 in the treatment group and n=34 in the control group. Mean changes from Stage 2 to Stage 3 were 0.60 days in the intervention group and 0.06 days in the control group. The standard deviations of the changes were 0.97 in the treatment group and 0.25 in the control group. As both the sample sizes and the standard deviations differed between the two groups, we used Hedge’s g to calculate the achieved effect size, which was 0.740. With an alpha of 0.05, the statistical power of the estimates for the between-subject comparison was 82.26%.

A power analysis for the between-subject effect that we had defined as relevant before conducting the experiment (i.e. a mean difference between the groups of 0.5 days) with an alpha of 0.05, a beta of 0.2, an SD of 0.65, and sample sizes of n=39 for the treatment group and n=34 for the control group, yielded a power of 85.06%.

Appendix A.7 – Post-Experimental Questionnaire

I. Socio-demographics

Your age: _____ years

Your gender: ☐ Male ☐ Female

What is your medical specialty? ________________________

Since when are you a consultant (specialist physician)? ____________________

When did you start practicing in the hospital? ____________________

II. Social and risk preferences

(‘Economic preferences’, according to Falk et al.[9,10] and Dohmen et al.[11])

1. How do you see yourself – Are you a person who is generally willing to take risks, or do you try to avoid taking risks? *Please indicate your answer on a scale from 0 to 10, where a 0 means “not at all willing to take risks”, and a 10 means “very willing to take risks”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

2. Please imagine that you have won a prize in a contest. Now you can choose between two different payment methods, either a lottery or a sure payment. If you choose the lottery there is a 50 percent chance that you would receive €1,000, and an equally high chance that you would receive nothing.
   What is the smallest sure payment that would make you prefer the sure payment over playing the lottery? Amount € _____________

3. How do you see yourself – Are you a person who is generally willing to give up some-thing today in order to benefit from that in the future, or are you not willing to do so? *Please use a scale from 0 to 10, where 0 means you are “completely unwilling to give up something today”" and a 10 means you are “very willing to give up something to-day”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

4. How well does the following statement describe you as a person? “I tend to postpone things even though it would be better to get them done right away.” *Please use a scale from 0 to 10, where 0 means “does not describe me at all” and a 10 means “describes me perfectly”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

5. How would you assess your willingness to trust strangers? *Please indicate your answer on a scale from 0 to 10, where a 0 means “not at all willing to trust strangers”, and a 10 means “very willing to trust strangers”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

6. How well does the following statement describe you as a person? “As long as I am not convinced otherwise, I assume that people have only the best intentions.” *Please use a scale from 0 to 10, where 0 means “does not describe me at all” and a 10 means “describes me perfectly”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

7. How do you assess your willingness to share with others without expecting anything in return when it comes to charity? *Please use a scale from 0 to 10, where 0 means you are “completely unwilling to share” and a 10 means you are “very willing to share”. You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

8. Imagine the following situation: You have won €1,000 in a lottery. Considering your current situation, how much would you donate to charity? (*Values between 0 and 1000 are allowed*): _____________

9. How well does the following statement describe you as a person? “When someone does me a favor I am willing to return it.” *Please use a scale from 0 to 10, where 0 means*

"does not describe me at all" and a 10 means "describes me perfectly". You can also use the values in-between to indicate where you fall on the scale.

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

10. How do you assess your willingness to return a favor to a stranger? *Please use a scale from 0 to 10, where 0 means you are "not willing to return a favor to a stranger" and a 10 means you are "very willing to return a favor to a stranger". You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

11. How do you see yourself – Are you a person who is generally willing to punish unfair behavior even if this is costly? *Please use a scale from 0 to 10, where 0 means you are "not at all willing to incur costs to punish unfair behavior" and a 10 means you are "very willing to incur costs to punish unfair behavior". You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

12. How well does the following statement describe you as a person? "If someone treats me unjustly, I will try to take revenge at the first occasion." *Please use a scale from 0 to 10, where 0 means "does not describe me at all" and a 10 means "describes me perfectly". You can also use the values in-between to indicate where you fall on the scale.*

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	1	2	3	4	5	6	7	8	9	10

III. Personality traits
(according to Gosling et al.[12] and Rammstedt and John[13])

In the following, you can find a number of personality traits that more or less apply to you. Please mark for each statement how well it describes your personality.

	Disagree strongly 1	Disagree 2	Somewhat disagree 3	Neither agree nor disagree 4	Somewhat agree 5	Agree 6	Agree strongly 7
1. I see myself as someone who is reserved.	☐	☐	☐	☐	☐	☐	☐
2. I see myself as someone who is generally trusting.	☐	☐	☐	☐	☐	☐	☐
3. I see myself as someone who tends to be lazy.	☐	☐	☐	☐	☐	☐	☐
4. I see myself as someone who is relaxed, handles stress well.	☐	☐	☐	☐	☐	☐	☐
5. I see myself as someone who has few artistic interests.	☐	☐	☐	☐	☐	☐	☐
6. I see myself as someone who is outgoing, sociable.	☐	☐	☐	☐	☐	☐	☐
7. I see myself as someone who tends to find fault with others.	☐	☐	☐	☐	☐	☐	☐
8. I see myself as someone who does a thorough job.	☐	☐	☐	☐	☐	☐	☐
9. I see myself as someone who gets nervous easily.	☐	☐	☐	☐	☐	☐	☐
10. I see myself as someone who has an active imagination.	☐	☐	☐	☐	☐	☐	☐
11. I see myself as someone who is considerate and kind to almost everyone.	☐	☐	☐	☐	☐	☐	☐

Appendix B – Model Specification

Our experimental data have a panel structure, as each pediatrician decided on the same randomly ordered 40 cases three times. Further, the decisions made in the experiment are nested within the subjects and the subjects are nested within the experimental sessions. To account for the hierarchical structure of our data with clustering on several levels, we applied multilevel mixed-effects panel regression models.[14] Our models include random effects for the experimental sessions, the subjects, and the 40 cases to account for potential within-group correlation of the decisions and for potential session-, subject-, or case-specific unobserved effects.

We employed the following model for a decision in experimental stage i on case j made by subject k, who is nested within session l (denoted by d_{ijkl}):

$$d_{ijkl} = \beta_0 + \beta_1 * \left(Stage2_{ijkl}\right) + \beta_2 * \left(Stage3_{ijkl}\right) + \beta_3 * \left(Treat_{kl}\right) + \beta_4 \left(Treat_{kl} * Stage2_{ijkl}\right) + \beta_5 \left(Treat_{kl} * Stage3_{ijkl}\right) + \beta_0^S * W_{kl}^S + \beta_1^S * W_{kl}^S * \left(Stage2_{ijkl}\right) + \beta_1^S * W_{kl}^S * \left(Stage3_{ijkl}\right) + \beta_1^M * X_{kl}^M * \left(Stage2_{ijkl}\right) + \beta_2^M * X_{kl}^M * \left(Stage3_{ijkl}\right) + u_{000l} + u_{00kl} + u_{10kl} * Stage2_{ijkl} + u_{20kl} * Stage3_{ijkl} + u_{0jkl} + \varepsilon_{ijkl}$$

The fixed-effects part of the model contains the constant β_0, fixed effects for Stages 2 and 3 of the experiment, which allow us to differentiate between the changes from Stage 1 to Stage 2 and the changes from Stage 2 to Stage 3, a treatment group indicator ($Treat_{kl}$), which is time-invariant, and two-way interactions between the treatment group indicator and the stage dummies. β_1 and β_2 denote the average changes over all subjects from Stage 1 to Stage 2 and from Stage 2 to Stage 3, respectively. β_3 is the average difference in the dependent variable between the treatment and the control groups, and β_4 and β_5 are average differences in changes over the stages between the two groups. Further, we included the subjects' individual characteristics in the fixed-effects part of our model. The vector W_{kl}^S (where $W_{kl}^{(1)}$ ... $W_{kl}^{(S)}$) contains S covariates, which are time-invariant characteristics of the individual subject k. We allow both the intercept and the changes between the experimental stages to vary at the subject level as a function of the subject characteristics S. The vector X_{kl}^M (where $X_{kl}^{(1)}$ ... $X_{kl}^{(M)}$) includes two-way interactions between the characteristics $m = \{1, \dots, M \leq S\}$ and the treatment-group indicator $Treat_{kl}$.[c]

[c] Note that X_{kl}^M does not stand alone but is either interacted with the Stage 2 or with the Stage 3 indicator. The reason is that we assume the interactions between the characteristics and the treatment-group indicator (denoted by X_{kl}^M) to be associated with the Stage 2 and the Stage 3 effects. In other words, while the effect of individual characteristics is assumed to be unassociated with the treatment group allocation in the first stage, it is in the second

The random effects are assumed to be independent of each other between levels and all random effects are independent of the level-one residuals. The residuals ε_{ijkl} are assumed to be independent and normally distributed with a mean of 0 and a constant variance σ^2 across the time points. Therefore, $\varepsilon_{ijkl}=\begin{pmatrix}\varepsilon_{0jkl}\\ \varepsilon_{1jkl}\\ \varepsilon_{2jkl}\end{pmatrix}\sim N(0,D)$, where $D=\begin{pmatrix}\sigma_1{}^2 & 0 & 0\\ 0 & \sigma_1{}^2 & 0\\ 0 & 0 & \sigma_1{}^2\end{pmatrix}$.

Further, we assume $u_{0jkl}\sim N(0,\sigma_2{}^2)$ for the random errors at the case level. The joint distribution of the three random effects associated with subject k (i.e., the random intercept denoted by u_{00kl} and the random slopes for Stage 2 and Stage 3 denoted by u_{10kl} and u_{20kl}, respectively) is $u_{kl}=\begin{pmatrix}u_{00kl}\\ u_{10kl}\\ u_{20kl}\end{pmatrix}\sim N(0,R)$. The random effects at the subject level u_{kl} are assumed to be multivariate normal with means of 0 and a variance-covariance matrix R, which is defined as

$$R=\begin{pmatrix}Var(u_{00kl}) & Cov(u_{00kl},u_{10kl}) & Cov(u_{00kl},u_{20kl})\\ Cov(u_{00kl},u_{10kl}) & Var(u_{10kl}) & Cov(u_{10kl},u_{20kl})\\ Cov(u_{00kl},u_{20kl}) & Cov(u_{10kl},u_{20kl}) & Var(u_{20kl})\end{pmatrix} \text{ or}$$

$$R=\begin{pmatrix}\sigma^2_{(3:intercept)} & \sigma_{(3:intercept,stage2)} & \sigma_{(3:intercept,stage3)}\\ \sigma_{(3:intercept,stage2)} & \sigma^2_{(3:stage2)} & \sigma_{(3:stage2,stage3)}\\ \sigma_{(3:intercept,stage3)} & \sigma_{(3:stage2,stage3)} & \sigma^2_{(3:stage3)}\end{pmatrix}.$$

We add the stage indicators to the random-effects specification at the subject level, as we are interested in the individual subjects' changes between the stages of the experiment. By including random slopes for the effect of the stages at the subject level, we allow for separate random effects within each subject for all stages. We allow correlation between the random effects at the subject level. The random effects at the session level are denoted by u_{000l} and assumed to be $u_{000l}\sim N(0,\sigma_4{}^2)$. We employ the same model specifications and assumptions for the analyses of the length of therapies and the appropriateness of therapy decisions. For regression results, see Table 3 and Table 5 in the main paper.

To analyze the association between pediatricians' individual characteristics and their antibiotic therapy decisions, we employed multilevel mixed-effects models. We used the same econometric model as described above without the panel time variables and the treatment-group indicator, as we considered only the decisions made in the first stage of the experiment when

---

and third stages where feedback was announced and given only in the treatment group. Therefore, we included the interaction between the characteristics and the effect of feedback only in the random slopes equations, but not in the random intercept equation at the subject level.

the instructions were the same for pediatricians in the control and the intervention group. For regression results, see Table 4 in the main paper.

Appendix C – Robustness Checks

We conducted several analyses to check the robustness of our main results. First, we analyzed the pediatricians' decisions before and after being given feedback on a case-by-case basis. Results of non-parametric statistical analyses support our main results. We found that, for the vast majority of the cases, the length of therapies decreased and the appropriateness of the length of therapies increased. In particular, we observed a decrease or no change in the therapy length for 37 out of the 40 cases, and a decrease or no change in the absolute deviation from the experts for 35 out of the 40 cases. Changes in the opposite direction for the remaining cases were not statistically significant (p>0.190 for number of days and p>0.196 for absolute deviation from the expert recommendations, Wilcoxon matched-pairs signed-rank tests); for a detailed analysis of the effect of feedback on a case by case basis, see Tables C.1 and C.2.

Table C.1: The effect of feedback on length of antibiotic therapies for each case

Cases (ordered by category)	The subjects' decisions on days of antibiotic therapy						Change in mean number of days	p-values
	Stage 2			Stage 3				
	mean	median	s.d.	mean	median	s.d.		
Neonatal infections								
1	1.46	0	2.35	1.00	0	1.75	-0.46	0.43
2	1.46	0	2.51	1.26	0	2.16	-0.21	0.32
3	4.10	5	3.57	3.62	5	2.88	-0.49	0.11
4	6.95	7	2.79	6.64	7	2.36	-0.31	0.31
5	5.90	5	3.37	6.13	5	3.61	0.23	0.22
6	9.64	10	3.78	8.72	7	3.00	-0.92	0.02
7	14.41	14	5.14	12.72	10	4.98	-1.69	0.00
8	3.05	3	3.68	2.82	3	2.83	-0.23	0.80
9	4.59	5	2.56	4.31	5	2.02	-0.28	0.68
10	7.62	7	3.70	6.10	7	2.23	-1.51	0.01
11	8.74	7	4.17	7.69	7	2.59	-1.05	0.13
12	10.62	10	4.83	9.67	7	4.24	-0.95	0.03
Infections of the CNS								
13	11.33	10	4.35	10.87	10	3.74	-0.46	0.77
14	15.18	14	4.07	14.23	14	3.77	-0.95	0.04
15	14.56	14	3.67	14.41	14	3.19	-0.15	0.95
16	13.85	14	4.69	13.64	14	4.31	-0.21	0.96

Bone and joint infections								
17	15.23	14	6.27	14.87	14	6.33	-0.36	0.21
Upper respiratory tract infections								
18	2.97	0	3.79	1.79	0	2.78	-1.18	0.00
19	7.44	7	3.22	6.69	7	1.91	-0.74	0.05
20	2.21	0	3.47	2.05	0	3.15	-0.15	0.65
21	2.05	0	3.53	2.26	0	3.53	0.21	0.19
22	5.21	7	3.25	4.69	5	3.33	-0.51	0.04
23	4.97	5	3.10	4.62	5	3.22	-0.36	0.16
24	8.49	7	3.14	7.82	7	2.21	-0.67	0.03
25	6.38	7	3.41	5.77	7	3.39	-0.62	0.09
26	7.97	7	3.75	7.49	7	2.63	-0.49	0.17
Urinary tract infections								
27	0.46	0	1.55	0.46	0	1.55	0.00	1.00
28	4.90	5	2.39	4.10	5	2.17	-0.79	0.02
29	8.46	7	3.48	7.77	7	2.76	-0.69	0.06
30	9.79	10	2.74	9.36	10	2.91	-0.44	0.11
31	12.03	10	6.37	10.90	10	6.00	-1.13	0.00
32	10.44	10	3.42	9.23	10	3.17	-1.21	0.02
Lower respiratory tract infections								
33	1.74	0	3.38	1.49	0	2.61	-0.26	0.98
34	5.38	7	3.03	5.54	7	2.97	0.15	0.97
35	11.54	10	3.95	10.82	10	3.58	-0.72	0.10
36	9.87	10	3.14	8.77	7	2.49	-1.10	0.01
37	8.46	7	2.01	8.00	7	1.95	-0.46	0.01
38	8.26	7	2.70	7.51	7	1.54	-0.74	0.02
39	14.90	14	5.54	14.03	14	5.18	-0.87	0.03
40	10.49	10	4.41	9.28	10	3.78	-1.21	0.01

Notes. This table shows the effect of feedback on length of antibiotic therapies at case level. It shows the average number of days subjects in the intervention group (n=39) chose prior to feedback (in Stage 2) and after feedback had been given (in Stage 3). p-values are shown for two-sided Wilcoxon matched-pairs signed-rank tests.

Table C.2: The effect of feedback on absolute deviation from the experts for each case

Cases (ordered by category)	The subjects' absolute deviation from the experts (in days)						Change in mean number of days	p-values
	Stage 2			Stage 3				
	mean	median	s.d.	mean	median	s.d.		
Neonatal infections								
1	1.68	1	1.82	1.28	1	1.22	-0.39	0.42
2	1.51	0	2.39	1.31	0	2.04	-0.20	0.32
3	3.20	3	2.28	2.71	3	1.50	-0.49	0.11
4	2.17	2	2.26	1.83	2	1.85	-0.33	0.10
5	3.61	2	2.80	3.84	3	3.04	0.23	0.51
6	3.08	3	3.55	2.44	2	2.55	-0.64	0.30
7	4.56	2	3.59	4.09	2	2.97	-0.47	0.13
8	2.65	2	3.13	2.42	2	2.16	-0.23	0.27
9	2.05	1	1.70	1.75	1	1.09	-0.30	0.28
10	2.78	2	3.39	1.61	2	1.75	-1.17	0.06
11	3.11	1	3.69	1.93	1	2.21	-1.18	0.02
12	4.10	3	4.13	3.31	2	3.50	-0.79	0.13
Infections of the CNS								
13	3.34	4	3.19	3.02	2.70	2.46	-0.32	0.77
14	6.83	6	4.07	6.02	5.65	3.53	-0.81	0.04
15	4.27	4	3.42	4.06	3.50	2.99	-0.21	0.95
16	3.99	2	3.12	3.60	2.15	2.93	-0.39	0.9
Bone and joint infections								
17	5.84	3.55	3.06	6.02	3.55	3.12	0.18	0.91
Upper respiratory tract infections								
18	3.18	2	2.43	2.41	2	1.35	-0.77	0.00
19	1.83	0	2.64	1.23	0	1.54	-0.60	0.17
20	2.66	1	2.36	2.50	1	1.99	-0.15	0.65
21	2.60	1	2.49	2.67	1	2.48	0.08	0.65
22	2.47	1	2.12	2.56	1	2.29	0.10	0.20
23	2.35	1	2.15	2.53	1	2.29	0.17	0.94
24	2.20	2	2.31	1.76	1	1.3	-0.43	0.15
25	2.92	2	2.19	2.82	2	1.99	-0.10	0.98
26	1.82	2	3.43	1.79	2	1.98	-0.03	0.51

Urinary tract infections								
27	0.74	0	1.37	0.74	0	1.37	0.00	1.00
28	2.14	2	1.74	1.91	2	1.16	-0.23	0.25
29	2.50	2	2.52	2.03	1	1.84	-0.47	0.07
30	2.50	2	1.75	2.49	2	1.74	-0.02	0.33
31	4.52	2	5.47	3.96	2	4.92	-0.56	0.78
32	2.74	2	2.74	2.39	2	2.13	-0.34	0.47
Lower respiratory tract infections								
33	1.91	0	3.10	1.66	0	2.31	-0.26	0.98
34	2.28	1	2.02	2.22	1	1.96	-0.06	0.41
35	3.28	2	2.18	3.13	2	2.04	-0.15	0.96
36	2.69	2	2.45	2.00	1	1.64	-0.69	0.01
37	1.71	1	1.09	1.58	1	1.11	-0.12	0.01
38	1.57	1	2.30	1.14	1	1.01	-0.43	0.33
39	4.04	3	4.01	3.75	3	3.58	-0.29	0.89
40	2.99	3	3.26	2.74	3	2.65	-0.26	0.66

Notes. This table shows the effect of feedback on absolute deviation from the expert recommendations at case level. For each case, the pediatricians' choices were compared to the experts' aggregate opinion for the respective case. It shows absolute differences between the pediatricians' choices and the expert recommendations prior to feedback (in Stage 2) and after feedback had been given (in Stage 3). Only the intervention group (n=39) is considered. p-values are shown for two-sided Wilcoxon matched-pairs signed-rank tests.

Second, instead of using multilevel mixed-effects panel regressions we ran ordinary least squares regression models. The estimation results were qualitatively and quantitatively very similar compared to those of multilevel mixed-effects model; see Table C.3.

Table C.3: OLS regressions on the effect of feedback on antibiotic therapy decisions

Dependent variable	Length of antibiotic therapy (in days)			Absolute deviation from the expert recommendations (in days)		
Model	(1)	(2)	(3)	(4)	(5)	(6)
Feedback	-0.048	-0.048	-0.526	-0.270	-0.270	-0.823**
(= 1 if intervention)	(0.758)	(0.760)	(0.660)	(0.480)	(0.481)	(0.371)
Second stage	-0.063	-0.063	-0.063	-0.086	-0.086	-0.086
(= 1 if second stage)	(0.107)	(0.107)	(0.107)	(0.077)	(0.077)	(0.077)
Third stage	-0.112	-0.112	-0.112	-0.085	-0.085	-0.085
(= 1 if third stage)	(0.108)	(0.108)	(0.108)	(0.068)	(0.068)	(0.068)
Effect of announcement	-0.082	-0.082	-0.082	-0.068	-0.068	-0.068
(Second stage x Feedback)	(0.146)	(0.146)	(0.147)	(0.118)	(0.118)	(0.118)
Effect of feedback	-0.633***	-0.633***	-0.633***	-0.397***	-0.397**	-0.397**
(Third stage x Feedback)	(0.197)	(0.198)	(0.198)	(0.150)	(0.150)	(0.150)

Individual characteristics						
Female (= 1 if female)			0.728**			0.228
			(0.364)			(0.223)
Experience (Years in hospital)			-0.114***			-0.076***
			(0.029)			(0.018)
Willingness to take risks			-0.251***			-0.045
			(0.081)			(0.041)
Extraversion			0.066			0.108
			(0.136)			(0.089)
Agreeableness			-0.013			-0.058
			(0.175)			(0.106)
Conscientiousness			-0.522***			-0.479***
			(0.188)			(0.101)
Neuroticism			0.066			0.112
			(0.113)			(0.084)
Openness			0.172			0.067
			(0.117)			(0.070)
Further individual characteristics (Economic preferences)	No	No	Yes	No	No	Yes
Case dummies	No	Yes	Yes	No	Yes	Yes
Constant	7.860***	1.600**	3.068**	3.033***	1.719***	3.594***
	(0.707)	(0.708)	(1.218)	(0.442)	(0.456)	(0.720)
Observations	8,760	8,760	8,760	8,760	8,760	8,760
Subjects	73	73	73	73	73	73
R^2	0.014	0.577	0.612	0.014	0.180	0.239

Notes. This table shows parameter estimates from OLS regressions. The interaction 'Third stage ×Feedback' indicates the effect of showing feedback to subjects. In Models (1) to (3), the dependent variable is 'length of antibiotic therapies (in days)'. In Models (4) to (6), the dependent variable is 'absolute deviation from the expert recommendations', measured in absolute values of the difference between the pediatricians' choices and the experts' recommended therapy length (in days). For each case, the subjects' choices were compared to the experts' aggregate opinion for the respective case. Robust standard errors, clustered at the individual-subject level, are shown in parentheses. 'Economic preferences' comprise validated measures for trust, reciprocity, and altruism, as well as time and risk preferences.[9-11] The variable 'case dummies', which is included in Models (2) to (3) and (5) to (6), indicates 40 dummies, one for each of the 40 medical cases. Furthermore, dummies for each experimental session were included in all models to control for any session effects. *** p < 0.01, ** p < 0.05, and * p < 0.10.

Third, we used two alternative measures for the pediatricians' deviation from the expert recommendations. Rather than the absolute deviation from a mean recommended length of therapy for each case, we used the absolute deviation from the interquartile range (IQR) of the expert decisions. Further, we analyzed how feedback affects the match between the experts' and the pediatricians' decisions. To this end, every decision on length of therapy from the experiment was replaced by the share of experts who chose exactly the same length of therapy for the particular case. The higher the share of experts who made the same decision, the larger was the match between the pediatrician's decision and the expert recommendations. Multilevel mixed-effects panel regressions with these outcome measures further corroborate our main findings regarding the effect of feedback on the appropriateness of care; see Models (1) and (2) in Table C.4.

Finally, we tested whether the changes in the pediatricians' decisions after provision of feedback are related to the difficulty of a case, measured in the case-specific heterogeneity in the experts' decisions. To this end, we calculated the standard deviation of the experts' recommendations on length of therapies for each case and applied a median split to form two categories: 'difficult to assess' and 'easy to assess'. We interacted our feedback variable with the indicator for the case category in order to analyze whether the effect of feedback was associated with the difficulty to decide on the appropriate length of therapy. The change in the number of days through feedback was not significantly affected by the difficulty of a case, while the change in absolute deviation from the experts was weakly significantly affected. For the latter, the effect of feedback was somewhat smaller for the hard cases; see Models (3) and (4) in Table C.4.

Table C.4: Robustness checks

Dependent variable:	Absolute deviation from IQR of the expert recommendations (in days)	Match with the expert recommendations	Length of antibiotic therapies (in days)	Absolute deviation from the expert recommendations (in days)
Model:	(1)	(2)	(3)	(4)
Fixed effects				
Feedback (= 1 if intervention)	0.712 (0.472)	-0.022 (0.028)	0.824 (0.522)	0.250 (0.364)
Second stage (= 1 if second stage)	-0.037 (0.100)	0.001 (0.005)	-0.063 (0.107)	-0.086 (0.088)
Third stage (= 1 if third stage)	-0.098 (0.140)	-0.002 (0.006)	-0.112 (0.150)	-0.085 (0.115)
Effect of announcement (Second stage x Feedback)	-0.095 (0.137)	0.008 (0.007)	-0.082 (0.147)	-0.068 (0.120)
Effect of feedback (Third stage x Feedback)	-0.567*** (0.192)	0.024*** (0.008)	-0.680*** (0.212)	-0.478*** (0.164)
Case category (=1 if hard to evaluate)			2.955*** (0.178)	1.161*** (0.085)
Case category x Effect of feedback (Third stage x Feedback)			-0.095 (0.112)	0.161* (0.090)
Constant	5.600*** (1.463)	0.290*** (0.060)	6.183*** (1.650)	3.963*** (0.840)
Random effects				
Session level				
Var(Constant)	0.143 (0.174)	0.001*** (0.001)	0.289 (0.283)	0.101*** (0.089)
Subject level				
Var(Stage 2)	0.202*** (0.057)	0.000*** (0.000)	0.226*** (0.066)	0.153*** (0.044)
Var(Stage 3)	0.530*** (0.112)	0.001*** (0.000)	0.595** (0.128)	0.342*** (0.076)
Var(Constant)	1.647* (0.469)	0.001*** (0.000)	1.638 (0.494)	1.103 (0.242)
Cov(Stage 2, Stage 3)	0.199*** (0.065)	0.000 (0.000)	0.232*** (0.075)	0.143*** (0.047)
Cov(Stage 2, Constant)	-0.278** (0.125)	0.000 (0.000)	-0.245* (0.137)	-0.286*** (0.084)

Cov(Stage 3, Constant)	-0.775***	0.000	-0.740***	-0.586***
	(0.051)	(0.000)	(0.215)	(0.124)
Case level				
Var(Constant)	18.666***	0.053***	21.887***	4.369***
	(0.519)	(0.001)	(0.610)	(0.136)
Var(Residual)	2.738***	0.013***	3.321***	2.191***
	(0.051)	(0.000)	(0.062)	(0.041)
Number of observations	8,760	8,760	8,760	8,760
Number of subjects	73	73	73	73
Number of sessions	8	8	8	8

Notes. This table shows parameter estimates from multilevel mixed-effects REML regressions. In Model (1), the dependent variable is ‘absolute deviation from the IQR of the expert recommendations (in days)’. The dependent variable in Model (2) is ‘match with the expert recommendations’, measured as the share of experts who made the same decision as the pediatricians in the experiment. Dependent variables in Models (3) and (4) are ‘length of antibiotic therapies (in days)’ and absolute deviation from the expert recommendations (in days)’, respectively. The interaction ‘Third stage × Feedback’ indicates the effect of showing feedback to subjects. The variable ‘case category’ in Models (3) and (4) is an indicator for the heterogeneity in the experts’ decisions (difficulty to evaluate the cases). Cases for which the standard deviation of the experts’ decisions on length of therapy was above the median were classified as cases that are ‘hard to evaluate’, while cases for which the standard deviation of chosen therapy durations was below the median, were classified as ‘easy to evaluate’. The interaction ‘Case category x Effect of feedback’ indicates the differential effect of feedback for easy and for hard cases. Standard errors are shown in parentheses. In all models, we control for the subjects’ gender, experience, Big Five personality traits,[12,13] and economic preferences, which comprise validated measures for trust, reciprocity, and altruism, as well as time and risk preferences.[9-11] All models include session-, subject-, and case-specific random effects. *** p<0.01, ** p<0.05, * p<0.1

References for Appendices

1. Deutsche Gesellschaft für Pädiatrische Infektiologie (DGPI). DGPI-Handbuch Infektionen bei Kindern und Jugendlichen. 7th ed. Stuttgart(Germany): Thieme; 2018.
2. Labenne M, Michaut F, Gouyon B, Ferdynus C, Gouyon JB. A population-based observational study of restrictive guidelines for antibiotic therapy in early-onset neonatal infections. Pediatr Infect Dis J. 2007;26(7):593–9.
3. Grol R, Dalhuijsen J, Thomas S, in't Veld C, Rutten G, Mokkink H. Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. BMJ. 1998;317(7162):858–61.
4. Grilli R, Lomas J. Evaluating the message: the relationship between compliance rate and the subject of a practice guideline. Med Care. 1994;32(3):202–13.
5. American Academy of Pediatrics (AAP). Chlamydia trachomatis. In: Kimberlin DW, Brady MT, Jackson MA, Long SS, editors. Red Book – 2015 Report of the Committee on Infectious Diseases. Elk. Grove Village (IL): Am Acad Pediatrics; 2015. p. 288–94.
6. Shah SS. Mycoplasma pneumoniae. In: Long SS, Pickering LK, Prober CG, editors. Principles and practice of pediatric infectious diseases. 4th ed. Edinburgh: Elsevier Saunders; 2012. p. 993–7.
7. Davey P, Marwick CA, Scott CL, Charani E, McNeil K, Brown E, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients. Cochrane Database Syst Rev. 2017;2, Art. No.: CD003543.
8. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175–91.
9. Falk A, Becker A, Dohmen T, Enke B, Huffman DB, Sunde U. Global evidence on economic preferences. Q J Econ. 2018;133(4):1645–92.
10. Falk A, Becker A, Dohmen T, Huffman DB, Sunde U. The preference survey module: a validated instrument for measuring risk, time, and social preferences. Institute for the Study of Labor (IZA). 2016; IZA Discussion Papers No.9674.

11. Dohmen T, Falk A, Huffman D, Sunde U, Schupp J, Wagner GG. Individual risk attitudes: measurement, determinants, and behavioral consequences. J Eur Econ Assoc. 2011;9(3):522–50.

12. Gosling SD, Rentfrow PJ, Swann Jr WB. A very brief measure of the Big Five personality domains. J Res Pers. 2003;37:(6):504–28.

13. Rammstedt B, John OP. Measuring personality in one minute or less: a 10-item short version of the Big Five inventory in English and German. J Res Pers. 2007;41(1):203–12.

14. Raudenbush SW, Bryk AS. Hierarchical linear models. Applications and data analysis methods. 2nd ed. Thousand Oaks(Calif.): Sage; 2002.