HSR Health Services Research


DOI: 10.1111/1475-6773.12040
SIMULATION METHODS IN HEALTH SERVICES RESEARCH: APPLICATIONS FOR POLICY, MANAGEMENT, AND PRACTICE

# Practice Variation, Bias, and Experiential Learning in Cesarean Delivery: A Data-Based System Dynamics Approach

*Navid Ghaffarzadegan, Andrew J. Epstein, and Erika G. Martin*

**Objectives.** To simulate physician-driven dynamics of delivery mode decisions (scheduled cesarean delivery [CD] vs. vaginal delivery [VD] vs. unplanned CD after labor), and to evaluate a behavioral theory of how experiential learning leads to emerging *bias* toward more CD and *practice variation* across obstetricians.

**Data Sources/Study Setting.** Hospital discharge data on deliveries performed by 300 randomly selected obstetricians in Florida who finished obstetrics residency and started practice after 1991.

**Study Design.** We develop a system dynamics simulation model of obstetricians' delivery mode decision based on the literature of experiential learning. We calibrate the model and investigate the extent to which the model replicates the data.

**Principal Findings.** Our learning-based simulation model replicates the empirical data, showing that physicians are more likely to schedule CD as they practice longer. Variation in CD rates is related to the way that physicians learn from outcomes of past decisions and accumulate experience.

**Conclusions.** The repetitive nature of medical decision making, learning from past practice, and accumulating experience can account for increases in CD decisions and practice variation across physicians. Policies aimed at improving medical decision making should account for providers' feedback-based learning mechanisms.

**Key Words.** Cesarean delivery, practice variation, experiential learning, simulation, system dynamics

Decision making regarding cesarean delivery (CD) versus vaginal delivery (VD) in the United States is suboptimal. First, there has been a *bias* toward more CD. Experts contend that CD is overperformed in the United States. In 2010, 32 percent of birth cases in United States were CDs (Hamilton, Martin, and Ventura 2011), whereas the World Health Organization's standard for developed countries is 10–15 percent (World Health Organization 1985). Second, there has been considerable *practice variation* in delivery mode

decisions, after controlling for patients' health status and risk. Epstein and Nicholson (2009) found substantial variation across both regions and physicians within regions, suggesting that physicians make different decisions for medically similar patients. Overperforming CD is undesirable for many reasons, including long-term side effects and increased costs (Hall and Bewley 1999; Wagner 2000; Villar et al. 2006).

There are many potential determinants of practice variation. Regional factors include culture, norms, standards, regulations, and the organization of health services (Fisher, Bynum, and Skinner 2009). Physicians' characteristics, such as age and training (Grytten and Sorensen 2003; Rebitzer, Rege, and Shepard 2007; Epstein and Nicholson 2009) and financial incentives also influence decisions (Bodenheimer and Grumbach 2005). Patients' characteristics such as health status, preferences, race, and socioeconomic status may influence physicians' decision making (Institute Of Medicine 2003). However, these factors cannot completely account for all practice variation and bias (Fisher, Bynum, and Skinner 2009). Accordingly, other relevant factors need to be considered.

Although the existing candidate explanations for practice variation come from a static "snapshot in time" framework, dynamic decision-making processes may be important. Delivery mode decisions are not one-time events. One obstetrician might perform thousands of deliveries in her career. One might expect that the repetitive nature of making decisions and observing outcomes (experiential learning) could influence practice style and lead to learning and correcting decisions. For example, obstetricians with similar training, background, and patient pools may by random chance perform different numbers of CD versus VD early in their careers. Those early initial experiences (including poor experiences such as emergency CD after initial labor) might change their tendencies to conduct CD versus VD, in turn affecting their skills and future performance of CD versus VD. This process of feedback and learning may lead to practice divergence despite identical initial states and patient populations. Such possible effects of experiential learning on practice styles are underexplored in the literature.

---

Address correspondence to Navid Ghaffarzadegan, Ph.D., M.B.A., Engineering Systems Division, Massachusetts Institute of Technology, Cambridge, MA; e-mail: navidg@mit.edu. Andrew J. Epstein, Ph.D., M.P.P., is with the Philadelphia Veterans Affairs Medical Center & Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA. Erika G. Martin, Ph.D., M.P.H., is with the Rockefeller College of Public Affairs and Policy and Nelson A. Rockefeller Institute of Government, University at Albany, State University of New York, Albany, NY.

This study aims at this direction and examines the contribution of experiential learning to practice variation. We develop a system dynamics model of delivery mode decision making (scheduled and unscheduled CD and VD) and explore dynamic changes in physicians' decisions as they accumulate experience. By imposing theory-based structures, a system dynamics simulation can use real-world data to test explanations that manifest endogenously, such as experiential learning. We calibrate our model with empirical data from deliveries performed by obstetricians in Florida and examine the extent to which simulation results replicate the data. We then analyze the simulation results and discuss policy implications.

## PRIOR LITERATURE

### *Experiential Learning and Suboptimal Medical Decisions*

Experiential learning studies focus on how people learn from outcomes of prior decisions. This theory may account for some physician practice variation and bias. Delivery is an example of a repetitive decision-making task. Physicians frequently make decisions, such as selecting scheduled or unscheduled CD or VD, enact decisions, and receive outcome feedback (such as maternal and child outcomes). Theoretically, experiential learning should improve decision making (Cyert and March 1963; Nelson and Winter 1982; Levitt and March 1988). However, information gathering and analysis are subjective, and learning from noisy, delayed, and conditional feedback is complicated (Sterman 1989; Huber 1991; Lant 1992; Levinthal and March 1993; Miner and Mezias 1996; Rahmandad 2008). A physician knows only about outcomes conditional on the chosen treatment; it is impossible to observe what the outcome would have been under an alternate treatment (Elwin et al. 2007; Stewart, Mumpower, and Holzworth 2012). There is an information asymmetry in delivery decisions because obstetricians can convert a VD into a CD during labor but not the reverse.

Although repetitive decision making has been studied in psychology, most studies have not examined effects of learning from outcome feedback on bias and variation in medicine. One exception is Ghaffarzadegan's (2011) theoretical model of suboptimal decisions in skill-sensitive tasks. The model shows that disagreement and bias can emerge endogenously through daily actions and outcome learning, even completely independently of external factors like financial incentives. When applied to the medical context, Ghaffarzadegan's (2011) theoretical model predicts that if patient outcomes

depend on physicians' skills, physicians will be more likely to select the procedures for which they have the highest skill. These decisions further improve their skills with the procedure, thereby increasing the future likelihood of selecting that procedure. However, that study does not use empirical data. This study refines the original theoretical model to control environmental factors, such as patients' and physicians' characteristics, health risks, and the secular trend, besides experiential learning, and uses a rich dataset to empirically test effects of learning on practice variation and bias in the context of obstetrics.

### *System Dynamics Modeling*

System dynamics is a simulation approach to analyzing dynamic problems arising in complex social systems (Forrester 1961; Sterman 2000). These systems are characterized by causal interdependence and circular causality in the form of feedback loops.[1] Model boundaries are large enough to endogenously capture a social system's behavior and identify system characteristics that generate observed behavior through feedback loops (Richardson 2011). This holistic approach results in models that are small enough to describe easily, while containing feedback loops that can replicate complex counterintuitive behaviors (Ghaffarzadegan, Lyneis, and Richardson 2011). Health-related examples of system dynamics approaches include models of polio eradication (Thompson and Duintjer Tebbens 2008; Rahmandad et al. 2011), chronic illness (Homer et al. 2010), U.S. health care reform (Milstein, Homer, and Hirsch 2010), tobacco use (Tobias, Cavana, and Bloomfield 2010), and the pharmaceutical market (Paich, Peck, and Valant 2011).

System dynamics can offer novel insights into the underlying processes of practice variation. Bias and practice variation are complex problems with different, but interconnected causal factors. Such complexity requires a dynamic feedback-based approach, which existing "snapshot in time" studies do not include. By incorporating a feedback-based structure, system dynamics models can test endogenous explanations for practice variation and bias. In addition, system dynamics models can predict potential effects of different policies.

## METHODS

### *Overview*

Our approach to evaluating obstetricians' delivery mode decisions involves two primary components, a system dynamics model and an empirical data

analysis. We formulate the system dynamics model by specifying theoretically based causal relationships in the delivery mode decisions. We calibrate the system dynamics model by analyzing Florida hospital discharge data on deliveries performed by 100 randomly selected obstetricians to estimate model parameters. We simulate the system dynamics model and generate a synthetic dataset based on how simulated physicians chose scheduled CD versus unscheduled CD versus VD throughout their careers. We validate the model by comparing the synthetic data to real-world data from two 100-physician samples not used in the model estimation. Finally we perform scenario analyses.

### *Data Source*

We used data from Florida all-payer hospital discharge databases from 1992 through 2008, covering all deliveries at nonfederal acute care hospitals. Florida data were selected because they contain physician identifiers in addition to patient characteristics and diagnosis and procedure codes. Cesarean deliveries were identified with an *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) procedure code of 74 in any procedure field. Vaginal deliveries were identified with ICD-9-CM diagnosis codes of 650 or 640.0*x* through 676.9*x* (where *x* is 1 or 2) in the principal diagnosis field and no indication of a cesarean delivery. The discharge data were augmented with information on each physician's medical training from the American Medical Association's Physician Masterfile. We identified "new" physicians as those who (1) completed their obstetrics residency training after 1991, (2) initially appeared in the discharge data in the same year as their residency completion, and (3) continued to perform deliveries in Florida through 2008.

Physician-level variables include the date they started practicing, gender, and medical school site (United States and Canada versus international). Time is reported in quarters of years. Delivery-level variables for patients include indicators for non-Hispanic white, Medicaid or no insurance coverage, previous CD, one or more of the 12 primary risk factors for CD identified by Gregory et al. (2002) (malpresentation, antepartum bleeding, herpes, eclampsia, uterine scarring, multiple gestation, macrosomia, unengaged fetal head, uterine abnormalities, other hypertension, preterm gestation, and congenital fetal anomalies), delivery mode (VD or CD), whether the patient labored, and major delivery complications, including laceration, hemorrhage, and infection (Srinivas et al. 2010). We apply a method developed by Henry et al. (1995) and Gregory et al. (2002) for using hospital discharge data to

determine whether a woman went into labor. The study was exempted from IRB review. The Appendix contains a descriptive summary of these variables.

We constructed three mutually exclusive samples of 100 randomly selected new obstetricians and their delivery decisions (VD, scheduled CD, and unplanned CD). Physicians in the first two samples finished their obstetrics residency and started practice between 1992 and 2000, whereas those in the third sample completed residency and started practice between 2001 and 2005. The first sample is used to calibrate the model and estimate parameters, whereas the other two samples are for out-of-sample comparisons of simulation results with data not used during calibration. In the first sample, 12 obstetricians are international medical graduates (compared to 8 and 18 in the second and third samples, respectively), 34 of obstetricians are female (compared to 43 and 68), and the average time of starting practice is 12 quarters after 1992 (compared to 26 and 47). In the first sample, 57 percent of patients are non-Hispanic white (compared to 57 and 45 percent), 45 percent were on Medicaid or had no insurance coverage (compared to 45 and 52 percent), 27 percent had a primary risk factor (compared to 27 and 32 percent), and 14 percent had a previous CD (compared to 14 and 14 percent).

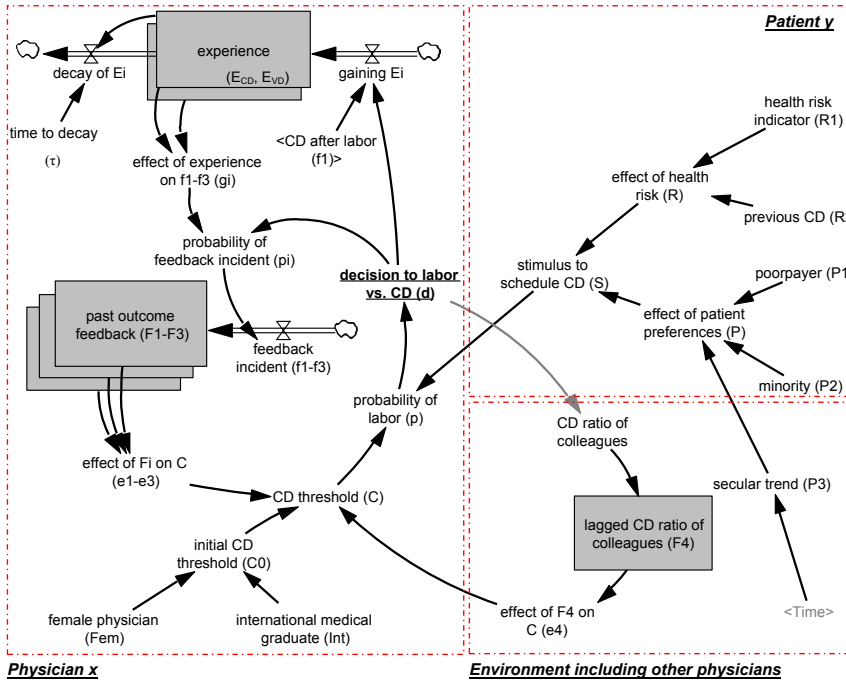
### *Simulation Model Description*

The system dynamics model simulates obstetricians' delivery mode decision making. Our unit of analysis is the physician, with discrete time units; in each time period, physicians meet patients, make delivery decisions, and learn from delivery outcomes. Our primary output is physicians' dynamic trend of CD and VD decisions. Figure 1 presents the model overview for *physician x* visiting *patient y*.

The model has three components. The left side is the main component, which includes physician decision making and learning. The final outcome is the delivery decision (scheduled CD vs. unplanned CD after labor vs. VD). The patient component is on the top right. This component's outcome is the *stimulus to perform CD (S)*, which is a combination of all information cues about patients' health risks and preferences. The bottom right represents other environmental factors that might influence physicians' decisions and patients' preferences, including colleagues and secular trends (specified as a linear time function).

Physicians select from three possible delivery modes: scheduled CD, unplanned CD after labor, and VD. Physicians first choose between scheduling a CD or labor; then for patients in labor they decide whether to revise the

Figure 1: A Simplified Conceptual Representation of the System Dynamics Simulation Model for Physician *x* Treating Patient *y*.




*Note*: The model has three components. The main component (feedback based) is the physician component. Arrows show the direction of causality. Variables in boxes represent stock variables, which are items that accumulate over time. Flows into and out of stock variables represent inflow and outflow variables, where in each time period $\Delta Stock = Inflow - Outflow$. The model includes 100 physicians and 251,800 simulated patients

decision to CD. The model assumes that physicians have unique thresholds, *CD threshold (C)*, representing their decision models for scheduling CD versus labor. To make delivery decisions, physicians compare their internal threshold with the information received from their patients, the *stimulus to schedule CD (S)*. If the stimulus is higher than the threshold, physicians schedule a CD. Over time, physicians accumulate experience and information about their past decisions (past outcome feedback). Physicians' thresholds can be affected by the *initial CD threshold* ($C_0$), a combination of their characteristics and background, and by past outcome feedback, shown by *effect of* $F_i$ *on C*.

We consider four types of feedback that might change physicians' *CD thresholds*. First, if the clinical risks increase while the patient is laboring, physicians can revise their initial decisions and perform unplanned nonelective CD. Observing the need to revise initial decisions is one feedback ($f_1$). Second, physicians might observe major complications. A complication during VD (which might not result in a revised decision) is another form of negative feedback ($f_2$). Third, physicians may observe major complications during scheduled CD ($f_3$), although it is not clear whether these major complications would be interpreted by physicians as negative feedback (CD complications are due to poor CD performance) or positive feedback (CD was necessary due to the patient's high-risk conditions, and outcomes would have been worse if VD were performed). These feedback incidents accumulate over time (past outcome feedback, $F_1$–$F_3$), representing lessons learned, and affect physicians' decision models. We construct $F_1$–$F_3$ by counting the total cumulative number of related incidents each physician faced over time since starting practice ($F_i = \sum f_i$). Finally, feedback can exist across colleagues, with physicians learning from each other. We have quarterly data on current and past delivery decisions from the same hospital where each physician works. The model randomly assigns physicians to different groups and formulates $F_4$ as a lagged average of CD percentage across simulated physicians within groups. We arbitrarily assume a group size of 5 physicians, based on group decision making studies that assume groups of 3–7 members (e.g., Hackman and Vidmar 1970; Cummings, Huber, and Arendt 1974), supplemented with informal conversations about our study with local physicians.[2] Physicians are influenced by the past decisions of their colleagues, and their decisions influence their colleagues' subsequent decisions. The effect size of these four types of feedback on CD threshold is estimated in the model calibration.[3]

The model also tracks CD and VD experience ($E_{CD}$ and $E_{VD}$). After a CD (elective or after labor), a physician gains one unit of CD experience, and after each labor, the physician receives one unit of VD experience. Experience decays with a delay (Huesch 2009). The feedback loops are closed by formulating each internal feedback incident ($f_1$–$f_3$), whose probability is expressed as a function of physician experience and patient health risk. For example, the probability of major complication during VD decreases as a physician's experience with VD relative to CD increases, and as the patient's health risk decreases.

The model in Vensim is available from the first author's website (NG). Consistent with Minimum Model Documentation Guidelines (Rahmandad

and Sterman 2012), Table 1 and Appendix SA5 contain all formulations and parameter estimates.

As Table 1 depicts, the *probability of labor* is a logistic function of *CD threshold (C)* minus *stimulus to schedule CD (S)*. *S* is operationalized as a function of patient *health risk* (whether the patient had a previous CD and at least one of the 12 indicators of risk) and *preferences* (whether the patient is a *minority* and has a *poor payer* insurance status, and *time* to control for environmental factors and the secular increase in CD demand). The CD threshold is operationalized as the net force of all past feedback: more negative feedbacks about VD will decrease the threshold, and more negative feedbacks about CD will increase the threshold.[4]

### *Model Calibration*

Parameter estimation was conducted via partial model calibration, whereby different pieces of the model are calibrated separately (Homer 2012). Partial model calibration is possible when there is an empirical dataset that allows focus on different parts of the model. It is particularly effective and mathematically efficient when the model formulation consists of additive functions. Partial calibration, in contrast to calibrating the whole model in one step, can identify how well different parts of the model replicate reality (Oliva 2003).

The model calibration is performed via four logistic regressions. The first regression is based on equations 1–9 in Table 1; the decision to labor is predicted by $R_1$, $R_2$, $P_1$, $P_2$, Fem, Int, $F_1$, $F_2$, $F_3$, and $F_4$ in addition to square of $F_1$, $F_2$, and $F_3$. In addition, we controlled for the secular trend ($P_3$) as a linear function of time. This regression estimates the parameters from *past outcome feedback* to *decision*. The next three logistic regressions are based on equations 23–30 in Table 1 and predict the path from *decision* to *feedback incident* ($f_1$–$f_3$). The dependent variables in these regressions are $f_1$, $f_2$, and $f_3$ and the independent variables are $E_{CD}$, $E_{VD}$, $R_1$, and $R_2$ in addition to the squares of $E_{CD}$ and $E_{VD}$.

### *Simulation Runs*

After calibration, the base case and scenarios were run to explore (1) the average CD trend over time, (2) the extent to which our model replicates empirical data on CD trends, and (3) the conditions that change practice variation and bias. The first item was assessed by simulating the base run and analyzing the average of CD trends of the 100 simulated physicians. The second was

Table 1: Variables in the System Dynamics Model and Formulation

| Notation | Description | Formulation |
|---|---|---|
| 1: p | Probability of labor, p($d_{VD}$ = 1) | $1/(1 + \exp(S - C))$ |
| 2: S | Stimulus to schedule CD | $R + P$ |
| 3: R | Effect of health risk | $\alpha_1 R_1 + \alpha_2 R_2$ |
| 4: P | Effect of patient preferences | $\alpha_3 P_1 + \alpha_4 P_2 + \alpha_5 P_3$ |
| 5: $P_3$ | Secular trend | Time in quarter |
| 6: C | CD threshold | $C_0 + e(F_1, F_2, F_3, F_4)$ |
| 7: $C_0$ | Initial CD threshold | $\alpha_0 + \alpha_6$ Fem + $\alpha_7$ Int |
| 8: e | Effect of feedback on C | $e_1(F_1) + e_2(F_2) + e_3(F_3) + e_4(F_4)$ |
| 9: $e_i$ | Effect of $F_i$ on C | $\alpha_{8,i} F_i + \alpha_{9,i} F_i^2$ |
| 10: $P_1$ | If patient has Medicaid or no insurance | pr($P_1$ = 1) = 0.45 (test 1 & 2), 0.52 (test 3) |
| 11: $P_2$ | If patient is a minority | pr($P_2$ = 1) = 0.57(test 1 & 2), 0.45 (test 3) |
| 12: $R_1$ | One of 12 health risk indicators | pr($R_1$ = 1) = 0.27 (test 1 & 2), 0.32 (test 3) |
| 13: $R_2$ | Previous CD | pr($R_2$ = 1) = 0.14 (test 1–3) |
| 14: Fem | Female physician | pr(Fem = 1) = 0.34 (test 1), 0.43 (test 2), 0.68 (test 3) |
| 15: Int | International medical graduate physician | pr(Int = 1) = 0.12(test 1), 0.08 (test 2), 0.18 (test 3) |
| 16: $F_1$ | All past unplanned CD incidents | $\sum f_1$ |
| 17: $F_2$ | All past major complications during labor | $\sum f_2$ |
| 18: $F_3$ | All past major complications during CD | $\sum f_3$ |
| 19: $f_1$ | An unplanned CD incident | pr($f_1$ = 1) = $p_1$ |
| 20: $f_2$ | A major complication incident during labor | pr($f_2$ = 1) = $p_2$ |
| 21: $f_3$ | A major complication incident during CD | pr($f_3$ = 1) = $p_3$ |
| 22: $F_4$ | Lagged CD ratio of colleagues | $(\overline{CD}_a + \overline{CD}_b + \overline{CD}_c + \overline{CD}_d + \overline{CD}_e)/5$ |
| 23: $p_1$ | Probability of $f_1$ = 1, if d = VD | $1/(1 + \exp(-g_1(R_1, R_2, E_{CD}, E_{VD}))$ |
| 24: $p_2$ | Probability of $f_2$ = 1, if d = VD | $1/(1 + \exp(-g_2(R_1, R_2, E_{CD}, E_{VD}))$ |
| 25: $p_3$ | Probability of $f_3$ = 1, if d = CD | $1/(1 + \exp(-g_3(R_1, R_2, E_{CD}, E_{VD}))$ |
| 26: $g_i$ | Effect of experience and health on $f_i$ | $\beta_{0,i} + \beta_{1,i} R_1 + \beta_{2,i} R_2 + \beta_{3,i} E_{CD} + \beta_{4,i} E_{CD}^2 + \beta_{5,i} E_{VD} + \beta_{6,i} E_{VD}^2$ |
| 27: $CD_x$ | CD ratio of *physician* ($\overline{CD}_x$ : lagged$CD_x$ | Average of CD decisions in a quarter |
| 28: $E_{CD}$ | CD experience | $\sum_{n=1}^{n=t} (d_{CD,n} - E_{CD,n-1}/\tau)$; *n* represents time periods |
| 29: $E_{VD}$ | VD experience | $\sum_{n=1}^{n=t} (d_{VD,n} - E_{VD,n-1}/\tau)$; *n* represents time periods |
| 30: $d_{CD}$ | CD decision (scheduled or unplanned) | $(1 - d_{VD}) + f_1$ |
| 31: $\tau$ | Average number of decisions in a quarter | 47 |

*Note.* In equations 10–15, test 1 is the base run to replicate the calibration dataset, and tests 2 and 3 are two out-of-sample tests. The values come from the datasets. Parameters (exogenous values that are constant across all agents and during the simulation) are presented by Greek letters ($\alpha$, $\beta$, and $\tau$). The values of parameters shown as $\alpha$ and $\beta$ are estimated in calibration and are listed in the Appendix SA5. The value of $\tau$ is estimated based on Huesch (2009, p. 1974). Variables that vary across time (such as CD Threshold) or agent (such as Female) are represented by English letters.

assessed by conducting one in-sample comparison (comparing simulation results with the empirical data used in the calibration exercise) and two out-of-sample comparisons (comparing simulation results with two samples not used in model calibration). One set of out-of-sample replications used data from younger physicians who started practicing after 2000, to better assess external validity. The third was addressed by three counterfactual scenarios (Zagonel et al. 2004). We examined the extent to which bias and variation appeared to result from experiential learning by turning off various feedback loops and analyzing simulated physicians' decisions in the absence of experiential learning. In all simulation tests, we ran our model of 100 physicians 1,000 times to reduce noise and enhance precision.

## RESULTS

### *Calibration Results*

The Appendix contains parameter estimates from the four logistic regressions used in the partial model calibration. The regressions showed statistically significant effects of past feedback ($F_1$–$F_4$) on decisions ($p$), and effects of experience ($E_{VD}$ and $E_{CD}$) on feedback incidents ($f_1$–$f_3$). In our first regression (dependent variable: decision to labor), the negative direction of the effects of $F_1$, $F_3$, and $F_4$ on decisions ($p < .01$) is consistent with what the theories suggest. The coefficients of the quadratic terms of $F_1$–$F_3$ are in the opposite direction of the coefficients of $F_1$–$F_3$, indicating a declining marginal effect of feedback as people receive more feedback (learning curve). The negative coefficient for $F_4$ shows that physicians are more likely to select CD if their colleagues recently performed more CD. We also checked whether experienced physicians treat more patients with a previous CD. The effect was minor; after 1,000 deliveries, the average probability increases by 1 percent.

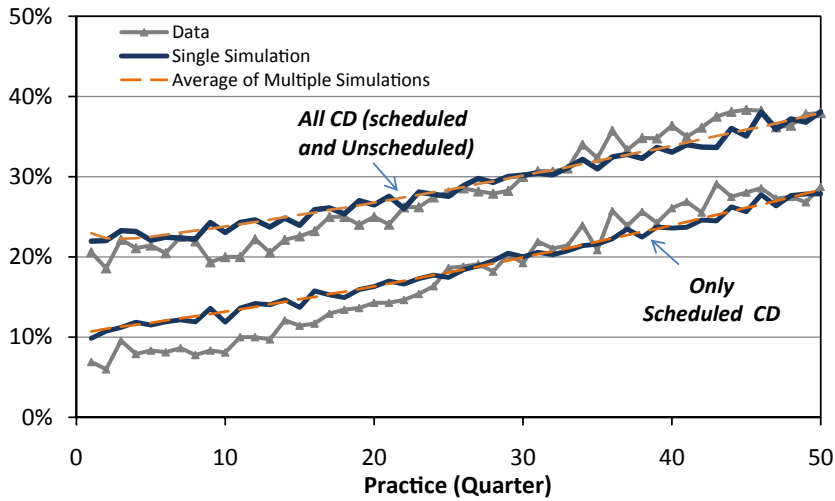
In the other logistic regressions (dependent variables: nonelective CD, labor with complication, and CD with complication), we found statistically significant effects of experience on feedback incidents ($p < .01$). The effect of experience on $f_1$ and $f_3$ is consistent with our expectation. On average, more VD experience (compared with CD experience) results in fewer nonelective CD ($f_1$), more major complication during VD ($f_2$), and more major complication during CD ($f_3$). The effect of VD experience on $f_2$ is negative at the beginning of practice, suggesting that when physicians are new, less VD experience results in more labor complications.

Estimated parameters were entered into the simulation model. First, we compared the average of CD tendency in 100 simulated physicians (a single simulation run, and average of multiple runs) with 100 real-world physicians in the calibration dataset (Figure 2). The average values from the simulations follow the average values of empirical data closely, with a correlation of 0.98. Physicians with more years of practice have an increased tendency to perform CD.[5]

*Model Validation*

Table 2 compares the simulation results with the empirical data for the percentage of scheduled CD and total CD. The first two columns (labeled "in-sample comparison") compare the results of simulation and calibration dataset. Between the first and fifth years of practice, simulated scheduled CD increases from 11 to 23 percent (in data, from 9 to 25 percent), and total CD increases from 22 to 32 percent (in data, from 22 to 35 percent). The next two sets of columns (labeled "out-of-sample comparison 1" and "out-of-sample

Figure 2: Simulation Base Run




*Note*: The graph shows the average probability of scheduling CD and performing CD (scheduled and unscheduled) over time. For each of these two variables, graphs present the probabilities from data (among 100 real physicians), single simulation run (100 simulated physicians), and average of multiple simulation runs (simulation of 100 physicians over 1,000 different runs).

Table 2: Comparison of Results of the Simulation Model with Data for Percentage of CD Decisions

| | *In-Sample Comparison* | | *Out-of-Sample Comparison 1* | | *Out-of-Sample Comparison 2* | |
|---|---|---|---|---|---|---|
| | *Simulation* | *Data* | *Simulation* | *Data* | *Simulation* | *Data* |
| During 1st year of practice | | | | | | |
| Sch. CD | 11% (5) | 9% (8) | 13%$^{+}$ (5) | 14%(14) | 19%$^{+}$ (6) | 20%(11) |
| All CD | 22%$^{+}$ (7) | 22% (13) | 25%$^{+}$ (7) | 24% (16) | 30%$^{+}$ (7) | 31% (14) |
| During 2nd and 3rd years of practice | | | | | | |
| Sch. CD | 13% (5) | 10% (7) | 16%$^{+}$ (6) | 15% (15) | 23%$^{+}$ (7) | 24% (8) |
| All CD | 23%$^{+}$ (7) | 22% (11) | 26%$^{+}$ (7) | 26% (8) | 34%$^{+}$ (8) | 35% (11) |
| During 4th and 5th years of practice | | | | | | |
| Sch. CD | 15%$^{+}$ (5) | 14% (9) | 19%$^{+}$ (6) | 20% (8) | 28%$^{+}$ (7) | 27% (9) |
| All CD | 26%$^{+}$ (7) | 25% (11) | 29%$^{+}$ (7) | 30% (10) | 38%$^{+}$ (8) | 38% (12) |
| After 5th year of practice | | | | | | |
| Sch. CD | 23%$^{+}$ (8) | 25% (13) | 28%$^{+}$ (8) | 26% (10) | 32%$^{+}$ (7) | 30% (15) |
| All CD | 32%$^{+}$ (8) | 35% (14) | 37%$^{+}$ (9) | 36% (12) | 42%$^{+}$ (8) | 40% (15) |

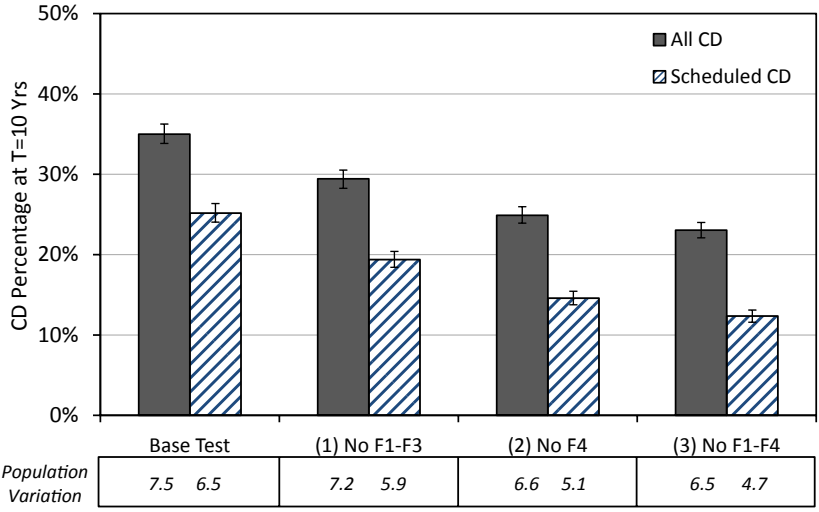
*Note. Sch. CD* stands for scheduled cesarean delivery, and *All CD* includes scheduled cesarean deliveries as well as unplanned cesarean deliveries after labor. Standard deviations in parentheses are in percentage points, and the ones under simulation columns show estimation of standard deviation across a population of 100 physicians. Simulation estimates with a plus sign show that there is no statistical difference between the model's prediction and the data ($p > .05$). The first out-of-sample comparison uses physicians who are drawn from a similar sample as those used in the model calibration and in-sample comparison with slightly different physician characteristics. The second out-of-sample comparison uses physicians who graduated more recently.

comparison 2") report our out-of-sample tests, validating the model by comparing the model behavior with datasets not used during calibration. The first out-of-sample comparison uses a random sample of physicians from the same generation, and the second out-of-sample test uses physicians who graduated more recently. Demographic characteristics of the first out-of-sample comparison group differ slightly from those of the calibration group due to sampling. The model can reasonably replicate the behavior of physicians not included in the calibration, including younger physicians. The model has a particularly close prediction in the first three time periods (year 1, years 2–3, years 4–5).

### *Counterfactual Analysis of Learning Effects among Physicians*

Finally, we tested effects of defined experiential learning mechanisms on CD decisions and compared them with the base test, in which all physicians start practice from the same initial conditions. Figure 3 reports the results of the simulation runs under three counterfactual situations for one cohort in their 10th year of practice if: (1) there was no learning from own decision outcome

Figure 3: What-If Counterfactual Analysis to Assess How the Percentage of CD Would Change under Three Scenarios of Reduced Learning




*Note*: The bar graphs show average results from 1,000 simulation runs for the population of 100 physicians with error bars showing the 95 percent confidence intervals of estimations. Population variation is measured as the average of the cohort-specific standard deviation across 1,000 runs. Base test is the case of 100 physicians starting from the same initial condition. In scenario (1) we turn off feedback loops that include the effect of F1–F3 in the model to examine *what if* there were no effects from one's past decision outcomes to decision model. In scenario (2) we turn off the effect of F4 to examine *what if* there were no feedbacks from others' decisions. In scenario (3) we investigate *what if* none of the four feedback mechanisms (F1–F4) work.

(excluding effects $F_1$–$F_3$ in the model), (2) there was no effect from CD trend in hospital (excluding effect $F_4$ in the model), and (3) there were no learning mechanisms (excluding effects $F_1$–$F_4$ in the model). Results are shown for both scheduled CD and all CD (scheduled and emergency). The first two bars show the percentage of deliveries that occur through CD in the base case, and the other sets of bars display outcomes under the three scenarios. The extent to which learning-based mechanisms explain bias can be assessed by comparing the first two and last two bars (Base Test vs. No $F_1$–$F_4$). By turning off the learning mechanisms, the scheduled CD rate changes from 25 to 12 percent (the total CD rate changes from 35 to 23 percent). The model's estimation of

within-cohort variation, measured as the standard deviation across physicians in a cohort, is reported at the bottom of the figure (population variation). By turning off the learning mechanisms, the variation across physicians in their scheduled CD rate decreases from 6.5 to 4.7 percent (interphysician variation in use of total CD decreases from 7.5 to 6.5 percent).

## DISCUSSION

Our simulation results suggest that experiential learning contributes to observed overuse (bias) and variation in scheduled CD decisions. The base run simulations show an increasing trend in the CD rate as physicians start practicing. The model controls for patient characteristics, health risks, whether patients had previous c-section, and the secular trend, and does not make any assumptions about physicians' financial incentives to perform CD. In the model, physicians' decision models may change as they accumulate experience and observe the results of their previous delivery decisions. These learning-based mechanisms partially replicate variation within a cohort, variation across different levels of experience, and the increasing trend of CD. The results provide evidence for the hypothesis that practice variation in and bias toward CD stem in part from the way that physicians learn from delivery outcomes.

### *Contributions*

This study demonstrates how experiential learning can contribute to practice variation and bias in medicine. Despite extensive research on the potential causes of practice variation, it cannot be explained entirely by measurable characteristics such as physicians' traits, financial incentives, region, and patient preferences. By connecting the problem of suboptimal medical decisions to the processes of skill development and learning from outcome feedback, our study differentiates itself from others that focus on financial systems and incentives (Bodenheimer and Grumbach 2005), patients' characteristics (Institute Of Medicine 2003), physicians' observable characteristics (Phelps 2000; Grytten and Sorensen 2003; Rebitzer, Rege, and Shepard 2007), and regional factors (Fisher, Bynum, and Skinner 2009) as main drivers of suboptimal decisions. Our approach to modeling delivery decisions allows for current decisions to affect subsequent ones dynamically and shows how similar obstetricians with similar patients could make different delivery

decisions after a few years. These findings suggest that even if policies to address the traditionally described causes of practice variation were implemented perfectly, variation would continue to occur as a result of these dynamic feedback processes.

This study makes several methodological contributions to both system dynamics and health services research. The study differs from snapshot-in-time studies of medical decision making (e.g., Way et al. 1998; Sorum et al. 2002). Our dynamic approach to study decision making shows how circular feedback processes can contribute to practice variation. In addition, we advance a prior theoretical model of medical practice (Ghaffarzadegan 2011) by using empirical data. Because health services data are so rich, future data-based system dynamics modeling could contribute to methodological developments in system dynamics, particularly around model calibration and testing.

### *Limitations*

This study has several limitations. Measurement error is unavoidable in administrative data. Our data are incomplete in measuring potentially important characteristics that influence delivery decisions, which is also reflected in our simulation model. For example, our data do not distinguish between induced versus unscheduled labor.

Some study limitations suggest avenues of future research. Our model captures dynamic mechanisms around the supply side of health services and for the specific case of delivery. This is a relatively simple decision (scheduled CD vs. unplanned CD vs. VD), and future work could evaluate more complex contexts and medical decisions. For example, additional feedback loops could capture the importance of patients' social networks in choosing physicians or preferred procedures and patients' decisions to see the same physicians for subsequent deliveries. Former patients may recommend specific physicians to their peers (Hoerger and Howard 1995), and physician groups may sort patients to specific physicians based on preferences or clinical needs (Epstein, Ketcham, and Nicholson 2010). Building on the literature of physicians' reputation formation (e.g., Navathe and David 2009), future work could examine the interactive effects of provider selection, reputation formation, and experiential learning. Finally, at the physician level more psychological theories can be incorporated. For example, the model can be developed to consider effects of depletion of past feedback, information overload, personal stress and emotions, and practice guidelines.

Other health conditions may require more sophisticated models to capture their more complex decision processes. For example, managing chronic mental illness includes decisions about the appropriate diagnosis, the selection of pharmaceuticals (including the offer of multiple psychotropic therapies), and drug dosing. Other chronic conditions such as HIV disease may require infectious disease clinicians to interact with other specialists for comorbid conditions such as substance use disorders or cancer, which may further complicate the decision-making process. We refrain from estimating the potential effects of specific interventions, such as checklists or training programs, due to inherent limitations in our data and modeling assumptions. We focus instead on insights into what interventions might work in the real world, and why. Future work could combine empirical data from actual interventions with simulation modeling to predict short- and long-term effects on practice variation.

### *Policy Implications*

We show that even if all determinants of practice variation that are typically discussed in the health services literature were addressed, some practice variation and bias would occur due to experiential learning. By demonstrating the effects of causal feedback loops, our study offers guidance to interventions to reduce practice variation. One implication from the strong effect of early-career experiences is that intervening during residency programs may have lasting effects, which is consistent with other work emphasizing the importance of residency training (Asch et al. 2009; Legnini 2011). However, our finding that CD rates increase with physicians' years of experience suggests that one-time interventions may not be as effective as repeated interventions.

Our findings about the effects of conditional feedback suggest that learning may differ if clinicians were provided with follow-up information on all patients. Having full information would increase the accuracy of physicians' perceptions of whether specific practices lead to better outcomes. Data sharing already exists in some clinical settings, including the Society for Thoracic Surgeons National Database, and the National Cancer Data Base. In addition, the strong effect of colleagues' feedback provides impetus for programs that match clinicians with outside practices for shared learning. For example, New York has organized quality improvement learning networks that allow HIV primary care providers to share clinical experiences and disseminate information. Similarly, there has been a recent interest in physician coaching services to help clinicians maintain their skills (Gawande 2011).

More broadly, this study illustrates how system dynamics modeling can be used by health services researchers to develop a deeper understanding of dynamic processes and provide guidance on the types of policies and interventions that are likely to have a positive impact. The emphasis on using feedback loops to generate behavior endogenously can yield important insights into *how* and *why* outcomes may change under different circumstances. In addition, the dynamic perspective makes system dynamics models naturally suitable for evaluating how short- and long-run effects differ. Combining system dynamics tools with rich public health datasets has the potential to make methodological contributions to system dynamics and health services research, in addition to improving program and policy design.

## ACKNOWLEDGMENT

*Joint Acknowledgment/Disclosure Statement*: The authors report no conflicts of interest.

*Disclosures*: None.

*Disclaimers*: None.

## NOTES

1. The word *causal* has different meanings across audiences and perspectives. System dynamics models are designed to allow variables to directly influence each other through feedback loops. System dynamicists focus on the physics of the relationships and deliberately introduce feedback loops. For example, a higher birth rate in an otherwise stable population would increase the population size, thereby generating more annual births. One explicit goal of system dynamics is to develop a deeper understanding of the potential effects of dynamic relationships.
2. In reality, physicians' selection of their colleagues is endogenous to the system, and network theories explain network formation. We simplify the process to focus on *learning from outcomes.*
3. Another type of feedback might exist that is arguably rare: physicians might judge their decisions to schedule for a CD unnecessary after follow-up visits.
4. Another approach is to define *CD threshold* as a stock variable, and *change in CD threshold* will be a function of each single feedback incidents. Although mathematically similar (ultimately, CD threshold = initial CD threshold + effect of all past feedbacks), our approach allows us to test a quadratic effect of past feedback on the threshold.
5. The simulation model is time discrete (Time step = 1 workday). However, we also checked simulation results with smaller time steps (0.5 and 0.25; compared with the

base case of 1). The overall results remain unchanged, indicating that the core modes of behavior are insensitive to the time step.

## REFERENCES


Asch, D. A., S. S. Nicholson, S. Srinivas, J. Herrin, and A. J. Epstein. 2009. "Evaluating Obstetrical Residency Programs Using Patient Outcomes." *Journal of the American Medical Association* 302 (12): 1277–83.

Bodenheimer, T. S., and K. Grumbach. 2005. *Understanding Health Policy: A Clinical Approach.* New York: Lange Medical Books/McGraw-Hill.

Cummings, L. L., G. P. Huber, and E. Arendt. 1974. "Effects of Size and Spatial Arrangements on Group Decision Making." *The Academy of Management Journal* 17 (3): 460–75.

Cyert, R. D., and J. G. March. 1963. *A Behavioral Theory of the Firm.* Englewood Cliffs, NJ: Prentice-Hall.

Elwin, E., P. Juslin, H. Olsson, and T. Enkvist. 2007. "Constructivist Coding: Learning From Selective Feedback." *Psychological Science* 18 (2): 105–10.

Epstein, A. J., J. D. Ketcham, and S. Nicholson. 2010. "Specialization and Matching in Professional Services Firms." *RAND Journal of Economics* 41 (4): 811–34.

Epstein, A. J., and S. Nicholson. 2009. "The Formation and Evolution of Physician Treatment Styles: An Application to Cesarean Sections." *Journal of Health Economics* 28 (6): 1126–40.

Fisher, E. S., J. P. Bynum, and J. S. Skinner. 2009. "Slowing the Growth of Health Care Costs – Lessons from Regional Variation." *New England Journal of Medicine* 360 (9): 849–52.

Forrester, J. W. 1961. *Industrial Dynamics.* Cambridge, MA: Productivity Press.

Gawande, A. 2011. "Personal Best: The Top Athletes and Singers Have Coaches. Should You?" New Yorker [accessed on October 3, 2011]. Available at http://www.newyorker.com/reporting/2011/10/03/111003fa_fact_gawande

Ghaffarzadegan, N. 2011. *Essays on Applications of Behavioral Decision Making in Public Management and Policy.* PhD Dissertation. Albany, NY: Department of Public Administration and Policy, University at Albany, State University of New York.

Ghaffarzadegan, N., J. Lyneis, and G. P. Richardson. 2011. How Small System Dynamics Models Can Help the Public Policy Process. *System Dynamics Review* 27 (1): 22–44.

Gregory, K. D., L. M. Korst, J. A. Gornbein, and L. D. Platt. 2002. Using Administrative Data to Identify Indications for Elective Primary Cesarean Delivery. *Health Services Research* 37 (5): 1387–401.

Grytten, J., and R. Sorensen. 2003. Practice Variation and Physician-Specific Effects. *Journal of Health Economics* 22 (3): 403–18.

Hackman, J. R., and N. J. Vidmar. 1970. Effects of Size and Task Type on Group Performance and Member Reactions. *Sociometry* 33 (1): 37–54.

Hall, M. H., and S. Bewley. 1999. "Maternal Mortality and Mode of Delivery." *Lancet* 354: 776.

Hamilton, B. E., J. A. Martin, and S. J. Ventura. 2011. *Births: Preliminary Data for 2010.* Hyattsville, MD: National Center for Health Statistics [accessed on October 1, 2012]. Available at http://www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr60_02.pdf

Henry, O. A., K. D. Gregory, C. J. Hobel, and L. D. Platt. 1995. "Using ICD-9 Codes to Identify Indications for Primary and Repeat Cesarean Sections: Agreement with Clinical Records." *American Journal of Public Health* 85 (8): 1143–5.

Hoerger, T. J., and L. Z. Howard. 1995. "Search Behavior and Choice of Physician in the Market for Prenatal Care." *Medical Care* 33 (4): 332–49.

Homer, J. B. 2012. "Partial-model Testing as a Validation Tool for System Dynamics (1983)." *System Dynamics Review* 28 (3): 281–94.

Homer, J. B., B. Milstein, K. Wile, J. Trogdon, P. Huang, D. Labarthe, and D. Orenstein. 2010. "Simulating and Evaluating Local Interventions to Improve Cardiovascular Health." *Preventing Chronic Disease* 7 (1): A18.

Huber, G. P. 1991. "Organizational Learning: The Contributing Processes and the Literatures." *Organization Science* 2 (1): 88–115.

Huesch, M. D. 2009. Learning by Doing, Scale Effects, or Neither? Cardiac Surgeons after Residency. *Health Service Research* 44 (6): 1960–82.

Institute Of Medicine. 2003. *Unequal Treatment: Confirming Racial and Ethnic Disparities in Healthcare.* Washington, DC: National Academies Press.

Lant, T. K. 1992. "Aspiration Level Updating: An Empirical Exploration." *Management Science* 38 (5): 623–44.

Legnini, M. W. 2011. "Can Low-Performing Hospitals Train High-Performing Residents?" *American Journal of Medical Quality* 26 (5): 408–10.

Levinthal, D. A., and J. G. March. 1993. "A Model of Adaptive Organizational Search." *Economic Behavior and Organization* 2 (4): 307–33.

Levitt, B., and J. G. March. 1988. "Organizational Learning." *Annual Review of Sociology* 14: 319–40.

Milstein, B., J. Homer, and G. Hirsch. 2010. "Analyzing National Health Reform Strategies with a Dynamic Simulation Model." *American Journal of Public Health* 100 (5): 811–9.

Miner, A. S., and S. J. Mezias. 1996. "Ugly Duckling No More: Pasts and Futures of Organizational Learning Research." *Organization Science* 7 (1): 88–99.

Navathe, A., and G. David. 2009. "The Formation of Peer Reputation among Physicians and Its Effect on Technology Adoption." *Journal of Human Capital* 3 (4): 289–322.

Nelson, R., and S. G. Winter. 1982. *An Evolutionary Theory of Economic Change.* Cambridge, MA: Harvard University Press.

Oliva, R. 2003. "Model Calibration as a Testing Strategy for System Dynamics Models." *European Journal of Operational Research* 151 (3): 552–68.

Paich, M., C. Peck, and J. Valant. 2011. "Pharmaceutical Market Dynamics and Strategic Planning: A System Dynamics Perspective." *System Dynamics Review* 27 (1): 47–63.

Phelps, C. E. 2000. "Information diffusion and best practice adoption." In *Handbook of Health Economics*, edited by A. J. Cuyler and J. P. Newhouse, pp 223–64. Amsterdam: Elsevier Science.

Rahmandad, H. 2008. "Effect of Delays on Complexity of Organizational Learning." *Management Science* 54 (7): 1297–312.

Rahmandad, H., and J. D. Sterman. 2012. "Reporting Guidelines for Simulation-Based Research in Social Sciences." *System Dynamics Review* 28 (4): 396–411.

Rahmandad, H., K. Hu, R. J. Duintjer Tebbens, and K. M. Thompson. 2011. "Development of an Individual-Based Model for Polioviruses: Implications of the Selection of Network Type and Outcome Metrics." *Epidemiology and Infection* 139 (6): 836–48.

Rebitzer, J. B., M. Rege, and C. Shepard. 2007. *Information Technology and Information Overload in Health Care.* Bingley, England: Economics Department, Weatherhead School, Case Western Reserve University.

Richardson, G. 2011. "Reflections on the Foundations of System Dynamics." *System Dynamics Review* 27 (3): 219–43.

Sorum, P. C., T. R. Stewart, E. Mullet, C. Gonzalez-Vallejo, J. Shim, G. Chasseigne, M. T. Sastre, and B. Grenier. 2002. "Does Choosing a Treatment Depend on Making a Diagnosis? US and French Physicians' Decision Making about Acute Otitis Media." *Medical Decision Making* 22 (5): 394–402.

Srinivas, S. K., A. J. Epstein, S. Nicholson, J. Herrin, and D. A. Asch. 2010. "Improvement in US Maternal Obstetrical Outcomes from 1992 to 2006." *Medical Care* 48: 487–93.

Sterman, J. D. 1989. "Modelling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment." *Management Science* 35 (3): 321–39.

———. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World.* Boston, MA: Irwin/McGraw-Hill.

Stewart, T. R., J. L. Mumpower, and R. J. Holzworth. 2012. "Learning to Make Selection and Detection Decisions: The Roles of Base Rate and Feedback." *Journal of Behavioral Decision Making* 25 (5): 522–33.

Thompson, K. M., and R. J. Duintjer Tebbens. 2008. "Using System Dynamics to Develop Policies that Matter: Global Management of Poliomyelitis and Beyond." *System Dynamics Review* 24 (4): 433–49.

Tobias, M., R. Y. Cavana, and A. Bloomfield. 2010. "Application of a System Dynamics Model to Inform Investment in Smoking Cessation Services in New Zealand." *American Journal of Public Health* 100 (7): 1274–81.

Villar, J., E. Valladares, D. Wojdyla, N. Zavaleta, G. Carroli, A. Velazco, A. Shah, L. Campodónico, V. Bataglia, A. Faundes, A. Langer, A. Narváez, A. Donner, M. Romero, S. Reynoso, K. S. de Pádua, D. Giordano, M. Kublickas, and A. Acosta. 2006. "Caesarean Delivery Rates and Pregnancy Outcomes: The 2005 WHO Global Survey on Maternal and Perinatal Health in Latin America." *Lancet* 367: 1819–29.

Wagner, M. 2000. "Choosing Caesarean Section." *Lancet* 356: 1677–80.

Way, B. B., M. H. Allen, J. L. Mumpower, T. R. Stewart, and S. M. Banks. 1998. "Interrater Agreement among Psychiatrist in Psychiatric Emergency Assessments." *American Journal of Psychiatry* 155 (10): 1423–8.

World Health Organization. 1985. "Appropriate Technology for Birth." *Lancet* 2: 436–7.

Zagonel, A. A., J. Rohrbaugh, G. P. Richardson, and D. F. Andersen. 2004. "Using Simulation Models to Address "What If" Questions about Welfare Reform." *Journal of Policy Analysis and Management* 23 (4): 890–901.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix SA2: Descriptive Statistics for the Key Variables in the Calibration Dataset.

Appendix SA3: Main Estimated Parameters and Standard Errors from the First Logistic Regression.

Appendix SA4: Main Estimated Parameters from the Second Set of Logistic Regressions.

Appendix SA5: Parameter Values for the Simulation Model and Source of Estimation.