Variety and Experience: Learning and Forgetting in the Use of Surgical Devices*

Kamalini Ramdas†, Khaled Saleh‡, Steven Stern§ and Haiyan Liu¶

November 2014

Abstract

A tremendous variety of medical devices is available to surgeons today. In this environment, a surgeon’s ease in using a device version that he has never previously used has important implications for cost and quality. Further, high device variety increases the time gap between repeat uses of any particular device version by a surgeon. This can result in forgetting over time of device-version-specific knowledge. While forgetting is inevitable, the impact of such forgetting over time at the level of specific tasks has not been examined previously. We use a unique, hand-collected dataset to examine learning and forgetting in hip replacement surgery as a function of a surgeon’s experience with specific surgical device versions and the time between their repeat uses. We also develop a generalizable method to correct for the left-censoring of device-version-specific experience variables that is a common problem in highly granular experience data, using MLE with simulation over unobservables conditional on observables. Even for experienced surgeons, the first use of certain component versions can result in about a 26% increase in surgery duration, hurting quality and cost. Also, with the passage of time, surgeons forget knowledge gained about the use of certain components. A three-month time gap between repeat uses of a component version in surgery can result in about a 50% drop in the time saving gained from past experience. We discuss implications for practice.

*We are grateful to Marianne Corbishley, Hyoun Ahn, Whitney Deck, and Amanda Wilson for help with data collection. For useful comments, we thank Dr. Wael Barsoum, Gerard Cachon, Sanjay Jain, Serguei Netessine, Wendy Novicoff, Nicos Savva, Elizabeth Teisberg, Karl Ulrich, and seminar attendees at Emory Business School, the Wharton Empirical Operations Conference, NYU and MIT. Mike Guthrie at Zimmer, Jerry Kie at Smith & Nephew, Chris Petrie at Stryker, and Robert McGlothlin at Depuy provided valuable industry insight and expertise.

†Management Science and Operations, London Business School, kramdas@london.edu.
‡Division of Orthopedic Surgery, Southern Illinois University School of Medicine, ksaleh@siumed.edu
§Department of Economics, University of Virginia, sns5r@virginia.edu.
¶Department of Economics, University of South Florida, hliu4@usf.edu.
1 Introduction

The past few decades have seen an explosion in variety in medical devices. In 2004, the US FDA\(^1\) was regulating 500,000 models of 1700 different medical devices (Maisel 2004).

Recent research in operations management and economics suggests that a proliferation in products, services, or tasks can slow down production due to limited learning spillover from one type of activity to the next (Benkard 2000, Ramdas and Randall 2008, KC and Staats 2012, Clark et al. 2012). Another potential reason for production slowdown in high variety settings is that it often can be a while before a worker performs any particular type of activity again, despite operating at high volume. This could lead to forgetting the intricacies of specific tasks. While learning curves have been estimated in many industries, the estimation of forgetting effects has received little attention outside of the laboratory (Bailey 1989). In particular, to our knowledge, no one has examined how forgetting with the passage of time at the granular level of specific tasks – such as a surgeon’s use of a specific surgical device – impacts production.

In this paper, we examine whether and how proliferation in surgical devices hurts production. In the surgical context, a slowdown in production amounts to a longer duration of surgery. All else equal, shorter duration is preferable as the risks of infection, blood loss, and post-surgical complexities are well-known to increase with surgery duration (Peersman et al. 2006, Yasunaga 2009). Also, longer surgery duration results in greater usage of expensive OR(Operating Room) capacity (Olivares et al. 2008, Saleh et al. 2009) and inefficient use of highly paid surgeons and other surgical staff. Yet, to our knowledge, no one has examined how learning or forgetting at the level of specific surgical devices impacts production.

We use a unique, hand-collected dataset from the University of Virginia Hospital to examine learning and forgetting at the level of specific surgical devices used in hip replacement surgery. We assembled data from multiple sources including OR records, patient charts, and hospital accounting databases. While many researchers have examined learning in surgery, we are the first, to the best of our knowledge, to examine how variety at the level of devices and, furthermore, individual device versions, impacts learning and forgetting. Our very detailed dataset enables us to consider a richer set of hypotheses than previously has been possible.

Productivity and quality reductions associated with initial learning on new device versions or with forgetting due to device proliferation may be viewed as a necessary evil if the long-term patient health benefits are superior to those that would have been obtained if a surgeon had used familiar device versions. Unfortunately, the current regulatory environment in many countries allows device vendors to sell devices that have no proven health benefit (Meier 2011, Kynaston-
Pearson et al. 2013), resulting in rapid introduction of device versions and a proliferation in available orthopedic devices (Gelberman et al. 2010). Accurate estimates of the short-term costs of such device proliferation are therefore crucially needed to inform policy targeted at improving patient outcomes and lowering healthcare provider costs.

In this paper, we estimate these costs of device proliferation in the context of hip replacement surgery. Our dataset includes all hip replacement surgeries performed at the University of Virginia Hospital from August 2006 until November 2008. We obtained information on the specific version of each of four key devices used in surgeries performed during this period – the "stem" or femoral device, which is inserted into the patient’s thigh bone; the "shell" or acetabular device, which is inserted into the patient’s hip socket; and the "head" and "liner" devices, which together comprise the ball and socket joint (see Figure 1). There are many versions of each of these four devices which differ in shape, material, coatings, and other characteristics that are likely to affect a surgeon’s ease in using them (see Figure 2 for two distinct stem versions). We limit our analysis to devices made by four vendors – Stryker, Depuy, Smith & Nephew, and Zimmer – that account for around 90% of the surgeries performed in our study period. When counting device versions made by only these four vendors, a total of 563 SKUs (or 121 unique device versions after accounting for devices that differ only in size) of these four key devices were used by just 4 surgeons in performing 671 hip replacement surgeries during our study period, indicating high variety in device versions.

Our dataset includes both first-time hip replacement surgeries as well as revision surgeries, which tend to take longer. Table 1 provides a simple description of our data, broken down into these two categories. Each of these categories is further divided into two subcategories. The first contains surgeries in which the surgeon had prior experience in our sample with the specific versions of all of the four key devices used in the surgery. The second contains the remaining surgeries in the category. A cursory glance at the mean values for surgery duration in each subcategory in the table suggests that surgeries that involve at least one device version that has not been used by a surgeon before in our sample period are of longer duration on average, both for revision surgeries and first-time surgeries. Of course, heterogeneity in patient characteristics, reasons for surgery, surgeon characteristics, and the operating room environment can confound this relationship, therefore, we will adequately control for these variables.

The impact of device proliferation should depend on how surgeons are trained. Traditionally, surgery has been taught using the apprenticeship model best exemplified by the phrase "see one, do one, teach one" (Gorman et al. 2000). This approach highlights the importance placed

\footnote{This table is based on a sample of 483 surgeries for which we have complete information on all variables used in our study. Details on how we assembled this dataset are presented in Section 3 below.}
by surgeons on a single unit of experience. While it is widely known that learning occurs steeply at first and then flattens out with experience (Wright 1936, Lieberman 1984, Argote and Eppe 1990, Argote 1999, etc.), in empirical estimation, emphasis is almost never placed on the very first unit of experience. Given the importance that traditionally has been placed on a single unit of experience in the surgical context, it is particularly useful to estimate the impact on surgery duration of a single previous exposure to a specific device version.

Importantly, since our data have a fixed starting time with no surgeries observed prior to it, we cannot be certain that an observed first usage of a specific device version by a surgeon is indeed the true first usage. In fact, even if we had data from when a surgeon joined the hospital, as surgeons typically get their initial training and experience at one hospital and then move elsewhere to practise, it would still be impossible to know if an observed first usage of a specific device by a surgeon is indeed a true first usage. In reality, given the tremendous and ever-changing variety in devices, the fragmented way in which device data are recorded, and the lack of attention that has been paid to this type of data in healthcare research and management to this date, it is very difficult to obtain current information on device usage at most hospitals, and even more so going back in time. This type of left-censoring of device usage data, while widely prevalent in hospital data, introduces a challenging econometric problem (Heckman and Singer 1985). A similar problem would arise with the use of highly granular experience data in other contexts. We present a generalizable approach to address this type of problem. In essence, we use observed information on the distribution of time between usages of specific devices to infer the probability that an observed first usage of a device by a surgeon is indeed a true first usage and incorporate this information into a maximum likelihood estimation procedure to estimate our coefficients of interest.

We find that a single prior usage of a stem (shell) version reduces surgery duration by about 26% (16%). Under conservative assumptions, this amounts to $32 Million in costs per year in the US due to stem and shell variety alone, excluding surgeon and anaesthetist costs. This time spent on first-time use of devices represents a potential 3.4% increase in the number of hip surgeries that can be performed annually. Forgetting also is costly. Halving the average time between repeat uses of stem and liner versions (by reducing the variety of these devices) would result in a saving of $13 Million in costs excluding savings in surgeon and anaesthetist costs per year in the US, representing a potential 1.1% increase in hip surgeries annually, under

---

3 An exception is Narayanan et al. (2009) who estimate the impact of software programmers’ first experience with a new software module.

4 An alternative is to use only surgeries where all devices used were introduced to the market after the starting time of the sample period. This approach can reduce sample size and result in non-randomly missing data. Also, it is impractical because hospitals do not store data on when specific device versions were introduced.
conservative assumptions.

Examining how device-specific learning and forgetting affects surgical outcomes requires a more granular level of analysis than procedure volume at an individual level. Our approach applies to other settings where the goal is to examine learning and forgetting at the level of subtasks or subprocesses of a service procedure. Client work in law, consulting, architecture and other professional services often entails combining elements from a variety of different knowledge bases, frameworks, or methodologies, although the speed of entry of new products may be significantly slower. Our approach is generalizable to such settings.

2 Hypotheses on Learning and Forgetting

Many researchers have documented the relationship between production volumes and both unit cost and quality (e.g., Wright 1936, Lieberman 1984, Argote and Epple 1990, Mukherjee et al. 1998, Argote 1999). In the field of healthcare, researchers in business, economics, and medicine have documented the impact of medical procedure volume on outcomes in a number of settings including many types of surgery (e.g., Birkmeyer et al. 2003, Huckman and Pisano 2006, Shwartz et al. 2008, Reagans et al. 2005, KC and Staats 2012, Clark et al. 2012, and Finks et al. 2011). Outcomes most commonly evaluated in this literature include mortality rate, the need for revision surgery, and procedure completion time or "duration of surgery." Several studies also have examined the impact of hospital volume and surgeon volume on outcomes in hip replacement surgery specifically. In a review of research on learning in hip replacement surgery, Shervin et al. (2007) find that hospital volumes and surgeon volumes are associated with improved outcomes. They call for further research to identify the causal factors – such as new surgical technology – underlying these volume-outcome relationships. Our dataset enables us to examine both the classical volume-outcome relationship at the level of individual surgeon experience as well as learning and forgetting with respect to a critical dimension of surgical technology – the key devices used in surgery.

Researchers also have started to examine learning at levels more granular than product or service volume. Benkard (2000) finds limited learning spillover in terms of production time in switching from production of one commercial aircraft model to another. Ramdas and Randall (2008) find limited learning spillovers for carmakers who use different brake components for different cars, Narayanan et al. (2009) for programmers who perform maintenance tasks on different software modules, KC and Staats (2012) for surgeons who perform different procedures for minimally invasive heart surgery, and Clark et al. (2012) for remote radiologists who read scans for different body organs or from different hospitals. We estimate the impact of experience
at the level of specific device versions on duration of surgery. In particular, we estimate the impact on duration of the very first use of a device version by a surgeon, an econometrically challenging task.

Naturally, one would expect considerable learning spillover from one task to another if the two tasks are quite similar. Relative to the differences among tasks examined in previous studies, one might expect the differences among versions of the same device, within the same type of surgical procedure, performed at a single hospital, to be smaller. Furthermore, in the case of orthopedic devices, it is widely acknowledged that most new devices are very minor variants of existing devices (Sheth et al. 2009, Demske 2008), unlike the case of two different aircraft models, brake designs, software modules, cardiac procedures, or body organs. The basic design of orthopedic devices has remained relatively stable for a few decades (Salemi, 2011, Gelberman et al. 2010, Bauer 1992). Manufacturers are well known to "tweak old models and patent the changes as new products" (Rosenthal 2013). In fact, provided a manufacturer can show that the new device is "substantially equivalent" to a legally marketed existing device, it can typically bypass clinical trials (Sheth et al. 2009) and go through a relatively straightforward patenting process that takes about six months, unlike the typical ten-year patent approval process in the pharmaceutical industry. In this environment, one would not expect a major penalty in terms of reduced productivity or quality for surgeries associated with first use of a "new" device variant.

In contrast to the vast literature on learning, little emphasis has been placed on knowledge depreciation (Argote 1999, Argote and Epple 1990). At the individual level, forgetting is a key cause of knowledge depreciation (Benkard, 2000). A critical determinant of the extent of forgetting is the amount of time between learning a task – such as how to use a specific device version – and recall of that learning the next time when it is needed (Wixted, 2004). When a limited number of tasks are being performed, the time gaps between repeated performance of any one task are likely to be smaller, and therefore the role of forgetting may be less important. On the other hand, when there is a large variety of tasks, as in our context, forgetting is more likely to come into play as any one task is performed less frequently.

Much of the literature on individual knowledge depreciation comes from the field of psychology and consists of theory and laboratory experiments (Bailey 1989), with little empirical estimation of individual knowledge depreciation rates outside of the laboratory. Keane and Wolpin (1997) estimate workers’ knowledge depreciation as a function of occupation switches between white collar and blue collar work and time out of work. Mincer and Ofek (1982) and Anderson et al. (2002) estimate depreciation in general human capital, using variation in time spent out of the workforce by working mothers. Yamaguchi (2012) treats a worker’s human
capital as a vector of skill levels associated with a variety of tasks, and models constant depre-
ciation of skills from one year to the next. As there are no gaps in the workers’ tenure in his
dataset, he cannot examine how the time since last performing a task impacts skill deprecia-
tion. In contrast, our dataset tracks the usage by individual surgeons of specific device versions
over time, including the time between repeat usages of each version, enabling us to estimate
knowledge depreciation at the level of specific tasks as a function of time since performing each
task. This approach provides a natural way to think about knowledge depreciation that is also
supported by research in psychology (Bailey 1989). To our knowledge, we are the first to be
able to estimate knowledge depreciation as a function of time at the level of individual tasks.

An underlying issue in knowledge accumulation and depreciation is the transferability of
what one has learned. For example, when one learns how to differentiate a polynomial, one does
not have to start over if the next polynomial to be differentiated has different coefficients or
different variable names. However, differentiating a different class of functions may involve some
separate learning and/or depreciation. Are different device versions like different polynomials
or are they like different function types, or are they like integration? If there really is something
to learn (and forget) this would suggest that they are more than just different polynomials with
different coefficients or variable names. Our analysis sheds light on this underlying question.

3 Data

We obtained data from the University of Virginia hospital for all hip replacement surgeries
performed from August 2006 to November 2008. Data on all devices used in each surgery
were obtained from a hospital database that is used for operational and accounting purposes.
Despite there being many studies of learning in surgery, to our knowledge no other study has
examined learning at the level of devices. This may be due in part to the difficulty in accessing
detailed data on device usage. We use our data to develop measures of surgeon experience
at the level of specific device versions. We supplement this data with data on outcome and
control variables from multiple sources including hand-collected data from individual patient
records, other hospital databases, and hand-collected data from records kept in the operating
theaters. Hand-collection of data was a painstaking process. We hired three nurses to perform
this task. Since a patient’s medical record was often a thick binder covering all visits to the
hospital and its associated clinics, finding and correctly interpreting the relevant data required
trained medical expertise. As an example, we needed to locate and read through the surgical
note for every patient in order to identify reasons for surgery and complexities during surgery.
Similarly, obtaining information from the records kept at operating theaters required our nurse
research assistants to access these paper documents through the operating theater nurses.

During our sample period, 752 hip replacement surgeries were performed by four surgeons at the University of Virginia hospital. Column (1) of Table 2 shows how these 752 surgeries are distributed across the four surgeons. These data are used to define the variable that measures the total experience for each surgeon during the sample period at the time of each surgery. Column (2) of Table 2 contains frequency by surgeon of surgeries in which at least one of the main devices used (head, stem, liner, and shell) were made by one of the four major vendors (Stryker, Depuy, Smith & Nephew, and Zimmer). This sample of 671 surgeries accounts for almost 90% of our sample. We chose to limit our sample in this way as we needed to interact closely with a vendor representative from each vendor to correctly classify component versions in our dataset. We use this sample to define our surgeon-specific component-version experience variables at the time of each surgery. Due to missing values on some control variables and in order to include only those surgeries for which all of the major devices were from one of the four major vendors, our sample is further reduced to 483 surgeries for which we have complete data. Column (3) contains frequency of surgeries by surgeon for this final sample that we use for our empirical analysis. We discuss below how we use these three levels of data to define the variables used in our estimation procedure. Columns (4) to (7) provide additional information about the education and professional background of all surgeons in our data sample. All our surgeons are highly experienced. We will discuss how this impacts our results in section 6.

Outcome Variable

Our outcome variable is \textit{duration of surgery}, defined as the amount of time in minutes from the start of a surgery, i.e. skin opening, until the end of the surgery, i.e. skin closing. Our measure of duration does not include the time taken to anesthetize the patient or the time that the patient may remain in the operating theater to "wake up" before being taken to the post-anesthesia care unit. We use the natural log of duration, resulting in the widely used log-linear experience curve (Reagans et al. 2005). As discussed in Section 2 above, duration of surgery is a commonly used outcome measure which impacts both quality and cost.

Experience Variables

We define a surgeon’s \textit{total experience} as the number of hip replacement surgeries that the surgeon has performed during the study period prior to a particular surgery considered. We

---

5 Two other surgeons performed 11 surgeries in total. We exclude these due to the low volumes.
6 We use the terms "device" and "component" interchangeably throughout.
7 For example, often the same component variant was recorded under slightly different names, needing an expert to identify the underlying variant.
calculate total experience for each surgeon using all 752 surgeries completed by the four surgeons in our sample.

Aside from gaining overall experience over time, each surgeon also accumulates experience over time with specific device versions. While many minor devices including screws and springs are used in each surgery, we learned from our discussions with orthopedic experts including our orthopedic surgeon coauthor that the shell, stem, liner, and head components are the primary drivers of the time taken to complete a surgery. These components also are quite expensive. The prices for components in our dataset ranged from $624 to $7,400 for shells, $1,525 to $6,955 for stems, $998 to $4,050 for liners, and $356 to $5,100 for heads. We focus on the possible learning and forgetting on these four main components.

The most granular level at which component experience can be accrued is the component SKU. A total of 114 unique shell SKUs, 162 unique stem SKUs, 122 unique liner SKUs, and 165 unique head SKUs were used in our sample period by just four surgeons, to perform 671 hip replacement surgeries which had at least one component from one of our four main vendors, as listed in Table 3. Within each of the four key components – shells, stems, liners and heads, component SKUs differ in technology, shape, materials, surface, coatings, and size.

For our purposes, we group together SKUs whose labels differ only in size into a single component version, for each of the four components and for each of the four vendors included in our study. In some cases, SKUs that differ only in size have slightly different item descriptions due to inconsistent use of abbreviations by the staff who originally recorded the data. Therefore, we enlisted the help of the hospital’s orthopedic device vendor representative for each vendor, to accomplish this grouping. Quite a bit of mixing and matching is possible over the versions of the four devices, both within and across vendors. Thus, it would be inappropriate to think of the appropriate unit of analysis as a fixed combination of specific device versions.

Through the procedure described above, the large number of SKUs was reduced to a much smaller number of component versions, as summarized in Table 3. In the sample of 671 surgeries, there are 20 shell versions, 35 stem versions, 28 liner versions, and 38 head versions, ignoring size variations. From now on, we use the term "component version" to denote all SKUs that vary only in size.

We created two types of surgeon experience variables at the level of specific component versions for each one of the four main components, by using the data from the 671 surgeries

---

8 For example, during the sample period, in the Zimmer line, the Trilogy Multi-Holed Shell component version contained 14 different size variations, ranging from 44mm to 70mm in diameter, while the Trilogy Uni-Holed Shell component version contained 11 different size variations, ranging from 46mm to 68mm in diameter.

9 These vendor representatives are present during surgery and have extensive knowledge of the specific device versions used for hip replacement surgery.
summarized in Column (2) of Table 2 as well as in Table 3. To our knowledge, neither of these variables has been examined in prior research on surgery.

For each surgeon, surgery, and component, the first time use dummy takes on the value of one if and only if the specific component version used in the surgery has not been used before by the surgeon in question during our study period.

For each surgeon, surgery, and component, experience gap is defined as the amount of calendar time (in days) since the last use by the surgeon of the specific component version used in the surgery. We use log values of "experience gap" variables to reduce the effect of outliers.  

Control Variables

We use a number of variables to control for the impact of patient and surgery characteristics on duration of surgery. Patient characteristics include age, gender, body mass index, anesthetic severity index, and patient comorbidities. Body mass index (BMI) is a standard measure of obesity of patients and calculated as the ratio of weight to squared height. BMI directly affects duration of surgery as a more obese patient can take longer to operate on. Anesthetic Severity Assessment (ASA) is another standard variable used in the medical literature that takes on integer values between 1 and 4 and is a rating of the overall fitness of the patient prior to surgery. The number of comorbidities is coded as the sum of ten indicator variables which indicate the presence of each of the ten most common patient comorbidities in hip surgery.

In addition, we control for a number of variables related to the surgery itself. Both Legs is a dummy indicating whether the surgery is performed on one or both hips. Surgeries that involve both hips generally are expected to take longer. Unihead is a dummy for the use of a unipolar head component. Through discussions with our orthopedic surgeon coauthor and other orthopedic experts, we learned that we can aggregate stem component versions from all vendors into two groups based on the method used for joining the component to the femur. Cemented stems have a smooth surface, and a cement-based adhesive is used to attach the stem to the femur. Uncemented stems, on the other hand, have a rough surface such that a proper

10 Note that we use log(1+experience gap) to avoid taking logs of zero.
11 At the UVA hospital, an ASA score for each patient is provided by both the anesthesiologist and a surgical team member. The two scores are highly correlated, and we use the average of the two scores. Our results are robust to the use of each of the individual scores.
12 The ten most common patient comorbidities are diabetes, kidney disease, liver disease, respiratory disorder, COPD, immune deficiency, prior venous thromboembolism, substance dependence, cardiovascular disease, high blood pressure, and bleeding disorders. In other specifications we also used the Charlson comorbidity index (Charlson 1987), a sum of indicators for presence of each five digit ICD9 code description for a patient condition, and a sum of indicators for presence of each three digit ICD9 code description (these are slightly more aggregate descriptions). We also estimated specifications in which we interacted complexity measures with the revision dummy and with time trend.
13 Unipolar heads are used in the treatment of hip fractures, which often involve a distorted anatomy and more bloody surgical field, resulting in longer duration.
joining of component and bone occurs when the bone grows around the component. *Cemented* is a dummy for use of a cemented stem. Revision surgeries do not always use all of the four main components. Therefore, we also include a dummy called *Use_component* for each of the four main components. For example, *Use_stem* is a dummy for use of a stem component.

We also control for the reasons for surgery. We include indicator variables for each of the most frequently cited reasons for surgery: Revision, Avascular Necrosis (AVN), Displasia, Arthritis, Severe Arthritis, End-stage Arthritis, Fracture, as well as an "Other Reasons" category that includes very infrequently cited reasons such as deformity, childhood disease, and post-traumatic bone conditions. The reasons-for-surgery dummies are non-exclusive, which means a surgery may have multiple reasons. For example, a surgery can be conducted because of arthritis and fracture. In the case of revision surgeries, we include an additional variable, "reasons for revision," which is the sum of indicator variables for each of the following reasons for revision surgeries: acetabular osteolysis, aseptic loosening, infection, pain, dislocation, and hematoma.

Finally, we include a linear and quadratic time trend\(^{14}\) to control for technological advances and other trends over time and surgeon-specific fixed effects to control for surgeon unobservables such as education and prior experience.\(^ {15}\) Table 4 reports the pairwise correlation coefficients of our experience variables. We do not find high correlation between our different experience measures.\(^ {16}\) Table 5 provides descriptive statistics of our main variables.

### 4 Empirical Specification

#### 4.1 Benchmark Specification

Prior research in business, economics, and medicine has found that surgeon-specific volume of surgeries performed is associated with better surgery outcomes and shorter surgery duration (e.g., Birkmeyer et al. 2003, Finks et al. 2011). We examine the impact of total experience on surgery duration using the benchmark specification,

\[
y_{st} = \beta X_{st} + \gamma e_{st} + u_{st} \tag{1}
\]

\(^ {14}\) Time trend is defined as the number of days since start of the sample period divided by 1000.

\(^ {15}\) Due to space constraints, some controls included in the regression are not reported in this table. Complete tables may be requested from the authors.

\(^ {16}\) We also compute eigenvalues of the inner product of explanatory variables and variance inflation factors for each of our variables. Both eigenvalues and VIFs are within acceptable ranges. All of our component-specific experience variables have VIFs below 2.8. The VIF values for all control variables are below 6.5, excluding total experience and the linear and quadratic time trend terms. Thus multi-collinearity is not a concern. The lowest eigenvalue is 0.02 and the corresponding condition index is 13.6 which is below the threshold value 30.
where $y_{st}$ is the log value of duration of the surgery performed by surgeon $s$ at time $t$, $e_{st}$ is the total experience of surgeon $s$ at time $t$, $X_{st}$ is a vector of control variables, and $u_{st}$ is the error term. Log-linear or "exponential" total experience curves are widely used in the literature to capture the diminishing returns from additional units of experience (Argote 1999, Thornton and Thompson 2001).

In prior research on learning in orthopedic surgery (e.g., Shervin et al. 2007, Yasanuga et al. 2009, Reagans et al. 2005), learning curves have been estimated without controlling for innate differences in the ability of individual surgeons. This can result in bias if more able surgeons attract more patients. Therefore, we modify the error term in (1) as $u_{st} = v_s + \varepsilon_{st}$, where $v_s$ is the individual surgeon fixed effect and $\varepsilon_{st}$ is the error term, to obtain the modified benchmark specification:

$$y_{st} = X_{st} + e_{st} + v_s + \varepsilon_{st}$$

Note that for each surgeon, experience accrued prior to our study period does not vary from one surgery to another within the study period and is fully captured via a surgeon dummy.

### 4.2 Component-Specific Experience

Next, we model the effect of component-specific experience over and above total volume of surgeries performed by a surgeon. We include our two measures of component-specific experience – whether each component version of the four main components has been used before and the time since prior usage of each of the four main component versions. As we control for total experience and include surgeon specific effects, the first measure enables us to estimate the extent of learning associated with a single usage of specific component versions. The second measure enables us to estimate individual surgeon forgetting over time at the level of specific component versions. We modify the benchmark specification as:

$$y_{st} = \beta X_{st} + \gamma e_{st} + \alpha w_{st1} + \theta \log [w_{st2}] + v_s + \varepsilon_{st}$$

where $w_{st1}$ and $w_{st2}$ are vectors of (observed) component-specific experience variables for surgeon $s$ at time $t$, as explained below.

Define $k_{st} = (k_{s1t}; k_{s2t}; k_{s3t}; k_{s4t})$ where $k_{sjt}$ is an index for the specific version of component $j$ used by surgeon $s$ in his $t^{th}$ surgery, where $j = 1, 2, 3, 4$ indexes the four main components – shell, stem, liner, and head. Next, we define $w_{st1} = (w_{s1t1}; w_{s2t1}; w_{s3t1}; w_{s4t1})$ where $w_{sjt1}$ is a dummy equal to 1 if and only if surgery $t$ is the first observed surgery using component $k_{sjt}$.

---

17 Note that we do not need a patient indicator because a surgeon operates on only one patient at a time.

18 We also use duration instead of its log value in alternative specifications, with qualitatively similar results.
Define \( w_{st2} = (w_{s1t2}; w_{s2t2}; w_{s3t2}; w_{s4t2}) \). For each \( sjt \) combination, \( w_{s1jt} \) is the observed time gap between the current surgery and the most recent prior surgery which used the component version \( k_{s1jt} \), if we observe a prior usage of component \( k_{s1jt} \) (i.e., if \( w_{s1jt1} = 0 \)). For those cases where we observe no prior usage of component \( k_{s1jt} \) (i.e., \( w_{s1jt1} = 1 \)), we set \( w_{s1jt2} = 0 \) without loss of generality. As mentioned previously, we use the logged value of \((w_{st2} + 1)\) in order to reduce the effect of outliers and to avoid taking the log of zero.

In the specifications in equations (2) and (3), we model an unobserved surgeon-specific fixed effect \( \nu_s \) and use surgeon dummies to capture this surgeon fixed effect.\(^{19}\) Since the error may have a different variance for each surgeon, we test for grouped heteroskedasticity using the test proposed by Levene (1960) and Brown and Forsythe (1974). The results shows that we cannot reject the null hypothesis; thus, we continue to use a homoskedasticity assumption.\(^{20}\)

Since we have an unbalanced panel with varying time gaps between observations for each surgeon (for example, a surgeon may do three surgeries on one day, none the next, and two the day after that), we construct a nonparametric estimator of the correlation between errors from two surgeries done by the same surgeon following Stern et al. (2010). In particular, we can write the error \( \varepsilon_{st} \) in any of the models we have discussed above as

\[
\varepsilon_{st} = \rho(d_{t,t-1}) \varepsilon_{st-1} + \eta_{st}; \eta_{st} \sim iid \left(0, \sigma^2_{\eta}\right).
\]  

(4)

The appendix shows how to estimate a correlation function \( \rho(d_{t,v}) \) for two surgeries \( t \) and \( v \) by the same surgeon as a function of the time gap between them, \( d_{t,v} = |t-v| \). Figure 3 displays the estimated correlation function \( \rho(d_{t,v}) \). We see that \( \hat{\rho}(0) \approx 0.25 \), implying that even for surgeries performed by a surgeon on the same day, the estimated correlation coefficient is relatively small. Also, the estimated correlation function dies out pretty quickly. Therefore, serial correlation in \( \varepsilon_{st} \) is not a concern even though we observe surgeons over a long time period.

We estimate specifications (1) through (3) using OLS. The results are presented in Table 6.

### 4.3 Correction for Left Censoring of Component-Specific Experience

A serious problem in the above specifications is that our two key measures of component-specific experience, namely, whether a specific component version is being used for the first time, \( w_{st1} \), and the amount of time since the last use of a specific component version by a
surgeon, $w_{st2}$, suffer from left censoring. This censoring problem arises for the first observed usage of component version $k_{sjt}$ by surgeon $s$: is it the true first, or was there a usage prior to the start of our sample? If there was a prior usage, then the true experience gap will be larger than the observed time gap between the start of our sample and the first observed usage of component $k_{sjt}$. Figure 4 illustrates this data censoring issue.\footnote{Note that the left censoring of the total experience of each surgeon, $e_{st}$, does not pose a problem as prior experience of each surgeon is fully captured by his/her surgeon fixed effect, $u_s$.} Clearly, this left censoring of component-specific experience is a serious problem as it affects our two main sets of variables of interest. This type of left censoring is a pervasive problem in hospital data because there is little data available on surgical component usage patterns going back in time, and also because surgeons typically move across hospitals over their careers. A similar concern would arise in other contexts when using highly granular experience data. Below, we develop a generalizable estimation procedure that corrects for this problem.

We rewrite specification (3) as

$$y_{st} = \beta X_{st} + \gamma e_{st} + \alpha z_{st1} + \theta \log [z_{st2}] + u_s + \varepsilon_{st} \tag{5}$$

where both $z_{st1}$ and $z_{st2}$ are vectors of (unobserved) component-specific experience variables for surgeon $s$ at time $t$. We define $z_{sjt1}$ as a dummy equal to 1 if and only if surgery $t$ is the true first surgery performed by surgeon $s$ using component $k_{sjt}$, and $z_{sjt2}$ as the true amount of calendar time since the last use of component $k_{sjt}$ by surgeon $s$.\footnote{If $z_{sjt1} = 1$, then we can just set $z_{sjt2} = 0$ (or any other constant) with no loss of generality.} For observations with $w_{sjt1} = 1$, $z_{sjt} = (z_{sjt1}, z_{sjt2})$ is not observed because surgeon $s$ may or may not have used the same component $k_{sjt}$ in a surgery prior to the beginning of the sample period. For notational simplicity, redefine $w_{sjt2}$ as the time gap between the starting day of our sample period and the date of the surgery at hand when $w_{sjt} = 1$. In fact, if $z_{sjt1} = 0$ \ $w_{sjt1} = 1$, then $z_{sjt2} > w_{sjt2}$ and therefore $z_{sjt2}$ is censored. For each component $j$, although we cannot observe the true values of $z_{sjt1}$ and $z_{sjt2}$ directly, if we can estimate the distribution of $z_{sjt} = (z_{sjt1}, z_{sjt2})$ conditional on observed $w_{sjt1}$ and $w_{sjt2}$, then we can completely overcome the censoring problem using simulation methods.

Recall that $z_{sjt2}$ is defined as the true amount of calendar time since the last use of the same version of component $j$ by surgeon $s$, therefore, it is a typical "time to event" variable and we can use survival analysis to deal with the data censoring issue. Let $S_j(\cdot)$ denote the survivor function of $z_{sjt2}$, i.e., the probability that $z_{sjt2}$ is larger than a certain value. By using the Kaplan-Meier estimator which takes into account the censoring problem, we can easily estimate $S_j(\cdot)$ nonparametrically from our data. Let $f_j(\cdot)$ be the density of $z_{sjt2}$ with distribution $F_j(\cdot)$. Then the estimate of $f_j(\cdot)$ and $F_j(\cdot)$ can be derived from $S_j(\cdot)$. We estimate the distribution
of \( z_{sjt2} \) for each component separately by assuming that their distributions are independent of one another and that each surgeon’s \( z_{sjt2} \) are drawn from the same distribution for all versions of component \( j \). See Figure 5 for a graph of the estimated \( F_j(\cdot) \) for each component.

Define \( h_j(z_{sjt} \mid w_{sjt}) \) as the density of \( z_{sjt} = (z_{sjt1}, z_{sjt2}) \) conditional on observed \( w_{sjt} = (w_{sjt1}, w_{sjt2}) \). We construct \( h_j(z_{sjt} \mid w_{sjt}) \) as follows. First, we define

\[
h_{j1}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 1, w_{sjt2}) = \begin{cases} 
F_j(w_{sjt2}) & \text{if } z_{sjt1} = 1, z_{sjt2} = 0 \\
f_j(z_{sjt2}) & \text{if } z_{sjt1} = 0; z_{sjt2} > w_{sjt2} 
\end{cases}
\]

Intuitively, the first line of equation (6) indicates that, when the observed first usage of a particular component version is the true first usage, \( z_{sjt1} = 1 \) and \( z_{sjt2} = 0 \). This case has a degenerate density \( F_j(\cdot) \).\(^{23}\) The second line indicates that, when the observed first usage of a particular component version is not the true first usage, then \( z_{sjt1} = 0 \) and \( z_{sjt2} \) must be larger than \( w_{sjt2} \). By the same logic, we have a degenerate density for \( z_{sjt} \) if we observe that there are prior usages of a particular component version by surgeon \( s \) before surgery \( t \). In this case, define

\[
h_{j0}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 0, w_{sjt2})
\]

with all of its mass at the values implied by \( w_{sjt2} \). Then we can write \( h_j(z_{sjt} \mid w_{sjt}) \) as

\[
h_j(z_{sjt1}, z_{sjt2} \mid w_{sjt1}, w_{sjt2}) = \begin{cases} 
h_{j1}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 1, w_{sjt2}) \\
h_{j0}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 0, w_{sjt2}) 
\end{cases}
\]

and define \( H_j(z_{sjt1}, z_{sjt2} \mid w_{sjt1}, w_{sjt2}) \) as the corresponding distribution function.

If we make a functional form assumption about \( \varepsilon_{st} \), then we can construct a likelihood term consistent with our imperfect knowledge of left-censored waiting time. In particular, we assume that:

\[
\varepsilon_{st} \sim iidN \left( 0, \sigma_{\varepsilon}^2 \right).
\]

Then, the likelihood contribution for \((y_{st} \mid X_{st}, e_{st}, z_{st}, v_s)\) is

\[
g(y_{st} \mid X_{st}, e_{st}, z_{st}, v_s) = \frac{1}{\sigma_\varepsilon} \phi \left( \frac{y_{st} - \beta X_{st} - \gamma e_{st} - \alpha z_{st1} - \theta \log[z_{st2}] - v_s}{\sigma_\varepsilon} \right).
\]

\(^{23}\)Note equation (6) is derived as

\[
\int h_{j1}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 1, w_{sjt2}) dz_{sjt} \\
= \int_{-\infty}^{w_{sjt2}} h_{j1}(1, 0 \mid w_{sjt1} = 1, w_{sjt2}) dz_{sjt2} + \int_{w_{sjt2}}^{\infty} h_{j1}(0, z_{sjt2} \mid w_{sjt1} = 1, w_{sjt2}) dz_{sjt2} \\
= F_j(w_{sjt2}) + 1 - F_j(w_{sjt2}) = 1.
\]
The likelihood contribution for surgeon $s$ performing surgery $t$ conditional on observed $w_{st} = (w_{st1}, w_{st2})$ can be obtained by integrating over unobserved variables, $z_{st}$, as\footnote{Since $z_{sijt}$ are independent from each other, the joint conditional density function of $z_{st} = (z_{s1t}, z_{s2t}, z_{s3t}, z_{s4t})$, $h(z_{st}|w_{st})$, can be written as $h(z_{st}|w_{st}) = \prod_{j=1}^{4} h_j(z_{sijt}|w_{sijt})$.}

\begin{equation}
L(y_{st} \mid X_{st}, e_{st}, w_{st}, v_s) = \int g(y_{st}, z_{st} \mid X_{st}, e_{st}, w_{st}, v_s) \, dz_{st}
\end{equation}

\begin{align*}
&= \int g(y_{st} \mid X_{st}, e_{st}, w_{st}, v_s) \prod_{j=1}^{4} h_j(z_{sijt}|w_{sijt}) \, dz_{st} \\
&= \int \frac{1}{\sigma_c} \phi\left(\frac{y_{st} - \beta x_{st} - \gamma e_{st} - \alpha z_{s1t} - \theta \log[z_{st2}] - v_s}{\sigma_c}\right) \prod_{j=1}^{4} dH_j(z_{sijt}|w_{sijt}).
\end{align*}

Then we can use simulation to approximate $L(y_{st} \mid X_{st}, e_{st}, w_{st}, v_s)$ (Stern 1997, Mcfadden 1989). The underlying intuition behind the simulation is to draw random values of $z_{sijt}$ from its distribution $h(z_{sijt}|w_{sijt})$ for each component $j$ and use them to compute the sample mean of

\begin{equation}
\bar{L}(y_{st} \mid X_{st}, e_{st}, w_{st}, v_s) = \frac{1}{R} \sum_{r=1}^{R} \left[ \frac{1}{\sigma_c} \phi\left(\frac{y_{st} - \beta x_{st} - \gamma e_{st} - \alpha z_{s1t} - \theta \log[z_{st2}] - v_s}{\sigma_c}\right)\right]
\end{equation}

is used to approximate (7).\footnote{Note that some of the components used by a particular surgeon are the same. One might think this causes $z_{sijt}$ to be dependent over $t$ if they share $k$. However, for this specification of the experience variables, there is no dependence because the randomness in $z_{sijt}$ applies only to the first observed occurrence of the use of type-$j$ component $k$.} Details of the simulation are provided in the appendix.

The likelihood contribution for all observations can be written as

\begin{equation}
L(y \mid X, e, w, v) = \prod_{s} \prod_{t} L(y_{st} \mid X_{st}, e_{st}, w_{st}, v_s).
\end{equation}

So Equation (9) can be simulated easily as

\begin{equation}
\tilde{L}(y \mid X, e, w, v) = \prod_{s} \prod_{t} \tilde{L}(y_{st} \mid X_{st}, e_{st}, w_{st}, v_s).
\end{equation}

Instead of choosing parameters to maximize the likelihood value of Equation (9), we choose parameters to maximize the simulated likelihood value of Equation (10). The results from MLE with simulation are presented in column (4) of Table 6.
4.4 Endogeneity

Surgery duration is chiefly determined by two factors: the complexity of the surgery itself and the "quality" of the lead surgeon who performs the surgery. The complexity of a surgery is determined by many factors including the reasons for the surgery, the physical condition of the patient, and whether it is a revision surgery. The "quality" of the surgeon is determined by his total experience, his component-specific experience, and other unobserved surgeon-specific attributes such as his training and innate dexterity or ability. Intuitively, all else equal, more complex surgeries should take longer to perform, while higher "quality" surgeons should finish faster.

In practice, at the UVA hospital, after being advised surgery, patients call into a call center and either ask for a particular surgeon, or, if not, are randomly assigned to an available surgeon. Thus patients may seek out surgeons with higher quality, causing high-quality surgeons to have more experience. This endogeneity problem is likely to cause a downward bias to the coefficient of total experience. On the other hand, if patients with more complex problems are more likely to seek out high quality surgeons, it is possible that the coefficient of total experience will be biased upward. Our inclusion of a large set of control variables for the complexity of a surgery, patient characteristics, as well as surgeon-specific fixed effects controls for the effect of these types of endogeneity. Note that prior research on the impact of experience on outcomes in hip replacement surgery has not included surgeon fixed effects, resulting in potentially biased results (e.g., Shervin et al. 2007, Yasanuga et al. 2009, Reagans et al. 2005).

Another important set of factors that can influence duration of surgery is the operating room environment at the time of any particular surgery. For example, a surgeon may try to speed up if the OR is highly congested and he is running late. However, surgeons decide which specific device versions to use in each surgery well in advance of the surgery date, in consultation with the patient. The surgery date is then fixed taking into account a number of factors such as patient availability and convenience as well as OR and surgeon availability. Thus, the level of congestion in the OR or how busy a surgeon’s schedule is on the surgery date will not impact choice of devices used in surgery. Therefore, not controlling for such variables will not bias our coefficients. Similarly, the composition of the surgical team in terms of specific nurses, attendants and others present may influence the duration of surgery, both directly and due to team experience effects (e.g., Reagans et al. 2005, Huckman et al. 2009, Huckman and Staats, 2011). Again, since the surgeon chooses the devices to be used in each surgery well in advance of the surgery date, device choice will be uncorrelated with composition of the surgical team for a surgery. Therefore omitting such variables will not bias our coefficients.
In selecting device versions for a specific surgery, a surgeon attempts to choose components that provide the best match with a patient’s specific needs. However, surgeons may prefer to use component versions that they are more familiar with for more complex surgeries, or to try out new component versions on simpler cases first. With the downward bias to the coefficient of first usage of a component version caused by this type of endogeneity, our results can be taken as "conservative."

5 Results

Benchmark Specifications

Column (1) of Table 6 contains results from our first benchmark specification. We see that total experience has a negative and statistically significant coefficient in the absence of any measures of component-specific experience. Other things equal, an increase of 100 surgeries in a surgeon’s total experience in hip replacement surgery will reduce the duration of surgery by approximately 11% on average. This result is consistent with past research on learning in hip replacement surgery which reports that the total experience of individual surgeons or hospitals can improve surgery performance and reduce surgery duration (e.g., Shervin et al. 2007, Yasanuga et al. 2009, Reagans et al. 2005).

Column (2) of Table 6 contains results for our second benchmark specification, which includes a surgeon dummy for each surgeon. Once we include these fixed effects, total experience is no longer significant. Thus, the negative coefficient of total experience in the specification in column (1) is likely due to variation in quality across surgeons – rather than due to within-surgeon learning with experience. Given the high experience level of our surgeons (see Table 2, columns (4)-(7)) it is not surprising to find that they appear to have reached the flat portion of their experience curves with regard to general learning about hip replacement surgery.26

We also see in both of these specifications that surgery duration decreases significantly with age of the patient, likely due to deterioration of muscle mass, which reduces the time taken to cut through muscle tissue. Surgeries on male patients take about 10% longer than those on female patients, likely due to differences in bone size and structure. Not surprisingly, revision surgeries take about 35% longer. In revision surgeries, cutting through scar tissue and removing previously implanted devices significantly adds to the time required for surgery. Surgeries in which both the left and right hip joints are operated on also display a significantly longer duration, though less than twice as long as a single side surgery. An increase in the number of

26We also ran a specification in which we allowed for a different coefficient for overall experience for surgeon 4, who had most recently completed his orthopedic fellowship. We see no effect for learning with overall experience.
comorbidities, which is a measure of the complexity of surgery, increases duration of surgery by approximately 4% for each additional comorbidity at a statistically significant level. Even after controlling for complexity of surgery in a variety of different ways\textsuperscript{27}, we observe a significant positive trend over time. Our discussions with surgeons at the UVA hospital indicate that, over time, there has been a trend at this hospital towards using specialized nurses outside their main area of focus in order to reduce the idle time costs associated with dedicated nurses. In fact, in a study of operating room efficiency conducted at the UVA hospital in the same timeframe as our study, McGowan et al. (2007) note that "nurses with specialized skills had drifted to various parts of the hospital, sometimes in places where the skills were not maximally used." The increased duration of surgery we observe over time is most likely an unintended negative consequence of this trend. Increased duration of surgery may also be due to greater usage over time of residents to complete more stages in the surgery, which typically adds to the amount of time involved. Based on discussion with surgeons, while they affect duration of surgery and costs, neither of these factors is likely to affect the choice of components used in a surgery.

\textbf{Component-Specific Experience}

We next consider component-specific experience at the highly granular level of component versions within each of the four key components. Column (3) of Table 6 includes our dummy variables for first usage of the versions used of the four main components as well as our continuous measures of time since last use of each of the component versions of these four main components. The model in column (3) is estimated using OLS, while the model in column (4) is estimated using our Maximum Likelihood Estimation procedure with simulation of unobservables conditional on observables. As described in section 4.3 we developed this procedure to overcome the left-censoring of our component-specific experience variables, which is a very pervasive problem in disaggregate data of this kind.

In column (3) of Table 6, we find that the first observed use of a stem version by a surgeon results in an approximately 23.8% increase in duration of surgery, all else equal, relative to cases where the surgeon has been observed using the stem version before. Note that, while we cannot be sure whether an observed first usage of a component version by a surgeon is indeed his true first usage of this version, the number of observed first usages can only exceed or at best equal the number of true first usages. If true first usages take longer than repeat usages, then the positive effect of true first usages will be watered down by including some later usages as first usages.

\textsuperscript{27}We included indicators for the ten most common patient comorbidites in THR surgery, and also used a variety of other measures as well as alternative specifications for complexity of surgery, in unreported regressions.
We also find in column (3) of Table 6 that there is knowledge depreciation over time in the case of both stems and liners. In the case of both stems and liners, a 1% increase in the number of days since previous usage of a specific component version results in an approximately 0.03% increase in surgery duration. When the experience gap for stems increases from its median (7 days) to its 75th percentile (24 days), surgery duration increases by about 3.1%, all else equal. Similarly, when the experience gap for liners increases from its median (9 days) to its 75th percentile (25 days), surgery duration increases around 2.6%, all else equal. Note that left censoring also affects the experience gap variable. However in this case, one cannot sign the bias. Therefore OLS coefficients cannot be considered as underestimates or overestimates of the true coefficients.

In column (4) of Table 6, we estimate the model using our modified MLE method with simulation to correct for left censoring of our key variables of interest. In accord with our intuition, we find that first usage of stems continues to result in a substantial (26% approximately) and highly statistically significant increase in duration of surgery with an increase in the size of the effect relative to OLS. Furthermore, we find that first usage of shell versions also results in a statistically significant increase in duration of surgery, albeit a smaller size of effect (16% approximately) than that for the first usage of stems. The significant effect of first use of shells on duration seen in column (4) suggests that this effect is masked in the case of OLS, where observed first use shells (which include both true first usages as well as cases where the shell version has in fact been used before) are used in estimating the effect.

In column (4), the coefficients of experience gap since last use of stems and liners continue to be statistically significant and of very similar magnitude to the OLS results. Also, in both specifications that contain component-specific experience variables, the coefficients for control variables are qualitatively similar to those in the benchmark specification.

Figure 6 further shows how both learning and forgetting effects for stems impact surgery duration, using results from column (4) of Table 6. Suppose a surgeon uses a new stem version for the first time some day. If he then uses the same stem version again that day, he will save almost 22.8% on surgery time relative to the first-use surgery, all else equal. However, if he waits for 10 days to use the same stem version, the time saving drops to 17.6%. Instead, if he uses it again after 3 months, the time saving drops to around 12.5%. In other words, we find that the time saving due to learning from past experience deteriorates with the time gap between two surgeries using the same stem version, due to forgetting over time.

\[28\] The interested reader may request a proof from the authors.

\[29\] Here, instead of using estimated coefficient directly to approximate the impact of learning and forgetting, we calculate the actual impact of past experience and forgetting by using the estimated parameters.
How an orthopedic device is implanted varies with the characteristics of the patient and the device. Patient characteristics include anatomy (e.g., body mass index, and presence of scar tissue), bone quality (e.g., presence of osteoporosis), bone shape (e.g., a thigh bone can have a narrow "champagne flute" canal with thick cortical walls, a moderate canal, or a wide "stove-pipe" canal with thin cortical walls – Dorr, 1989), and other comorbidities. Device characteristics include the geometry of the device, its length and thickness, materials, and type of porous coating. For stems, geometry includes the width of the flare at the top of the stem, the angle of the stem, and the "offset" or distance from the center of rotation of the head to the long axis of the thigh bone.

The placement of a stem into the thigh bone is the hardest part of a hip replacement surgery. It involves "broaching", i.e., using a mold to prepare and shape the bone canal, and "reaming", i.e., using a long thin drill to open it up. Slight differences in the shape of a stem can result in differences in how the stem is inserted because differently shaped stems can get caught up in different parts of the bone cavity. This results in the need for different surgical instruments for each stem version. The peculiarities of a new instrument are often only fully understood by using it. The variation in patient and device characteristics creates a significant amount of uncertainty and necessitates significant learning for stems, while also increasing the chances of forgetting over time.

Aside from needing to learn how to use a new instrument, there also can be particular idiosyncrasies associated with specific device versions. For example, we learned from one orthopedic surgeon that certain stems tend to sit a couple of millimeters higher when placed in the thigh bone canal, than the stem height specified on the outside of the box. A surgeon who does not know this would realize that a deeper opening is needed only after broaching, reaming and inserting the stem - causing all of these steps to be redone. This type of problem can be mitigated easily by highlighting such idiosyncrasies to surgeons new to a device.

Shell insertion is relatively easier than stem insertion for several reasons. The hip socket is fairly straightforward in terms of technical preparation. The surgeon needs to prepare the bony bed of the hip socket by reaming with a reamer that is hemispheric in shape. All shell versions have the same type of reamers and cup geometry, unlike the situation for stems.

Liners require delicate maneuvering as the mechanism for locking the liner into the shell is one of the most demanding parts of the surgery, although it does not take much time. Liners are tricky to insert because there is significant variation in the locking mechanism and also because there sometimes can be obstructive tissue or bone that requires to be dealt with in a slightly different way depending on the specific liner version used. The head is the most straightforward
component and does not require any instrumentation.

Our finding of significantly higher surgery duration in the case of surgeries involving first use of stem and shell versions is consistent with the notion that experience would be more significant for those components that require greater skill and dexterity to place properly. While first use of a liner version does not significantly impact surgery duration, this may be due to the fact that the total time needed to insert a liner is a small part of the total surgery time, even for a difficult insertion instance. It is not surprising that there is little learning or forgetting in the case of heads, which are easy to insert.

Significant depreciation of knowledge over time in the case of stems and liners is also consistent with the difficulty associated with these components. In the case of shells, we see no knowledge depreciation over time. This could be because most of the first-time learning is associated with idiosyncratic characteristics of specific shell versions, which, once learned, are unlikely to be forgotten.

6 Discussion and Conclusions

We have found that first-time use of a new stem version increases duration of surgery by about 26% (p-value 0.01). First-time use of a new shell version increases duration of surgery by 16% (p-value 0.10). These increases in duration proportionately increase the likelihood of infection, blood loss, and other complications. In our sample period, about 10% of all surgeries were first observed usages of a stem version, and 5% of all surgeries were first observed usages of a shell version. The average surgery duration in our sample is 165 minutes. With about 330 hip replacement surgeries performed each year in our sample period, this translates into 1219 additional minutes each year for surgeries involving first-time stems, and 398 additional minutes each year for surgeries involving first-time shells. Using a very conservative cost estimate of $20 per minute of OR time that is relevant for basic surgical procedures excluding surgeon and anaesthetist costs (Macario 2010), this amounts to almost $36,352 a year for just these two components used in hip replacement surgery, at the UVA hospital. In the US, 332,000 hip replacement surgeries were performed in the year 2010. Assuming similar rates of usage of unfamiliar devices as in the UVA hospital, this amounts to an additional $36.6 million in costs for stems and shells alone, excluding surgeon and anaesthetist costs. Assuming there is demand for surgery, freeing up time by reducing duration would allow hospitals to perform more surgeries. For example, at the UVA hospital, given average surgery duration of 165 minutes,

30Source: http://www.niams.nih.gov/Health_Info/Hip_Replacement/
ten additional hip replacement surgeries could have been performed per year in the additional time spent when operating with new stems or shells. This represents a potential 3.4% increase in the number of hip surgeries performed per year at the UVA hospital, and also at the national level under similar assumptions as above.

The above estimates of the costs of first use stemming from high device variety are conservative for two reasons. First, having access to data for only experienced surgeons has allowed us to examine the impact of high device proliferation on a highly experienced surgeon pool. The effects we observe would likely be even higher in the case of less experienced surgeons. Second, the OR is one of the most highly utilized suites in any hospital, and ORs are used for a variety of other orthopedic procedures, and also for many other procedures that involve a variety of devices and instruments. Variety in stems and shells is only a small but illustrative slice of the plethora of variety in devices in general (Maisel 2004).

Our estimated costs of forgetting are also high. There are two ways to think about the costs of forgetting – the cost due to the forgetting itself and the the cost due to high device variety. To estimate the former, note that if there is no forgetting, the coefficient $\theta$ in specification (5) is zero. The hypothetical surgery duration in this "no forgetting" case gives us a benchmark against which to compare the real surgery duration when forgetting occurs.\footnote{In our sample, there are 373 surgeries in which surgeons use a stem version which they have some previous experience with. For each of those surgeries, we use the following two equations to calculate the hypothetical surgery duration:}

$$\log(\text{duration}_{\text{real}}) = x + 0.027 \times \log(\text{gap}_{\text{stem}} + 1)$$

$$\log(\text{duration}_{\text{hypothetical}}) = x + 0 \times \log(\text{gap}_{\text{stem}} + 1)$$

where $x$ is everything else in the ln(duration) equation and its value remains the same under the hypothetical assumption, and and $\text{gap}_{\text{stem}}$ is the time gap between two surgeries where the surgeon used the same stem.\footnote{Some revision surgeries may not use a stem.}

In our 28-month sample period, about 3523 minutes in total are added to surgery duration due to forgetting related to stems, i.e. 1510 extra minutes a year. Similarly, for liners, the total time added due to forgetting is about 1410 extra minutes a year. Using similar estimates of per minute costs and total volumes as above, in the US, forgetting on stems and liners alone results in about $58.7 million in cost excluding surgeon and anaesthetist costs, or an opportunity loss of 5.3% more hip surgeries that could be performed, annually.

Although discussing the cost of forgetting itself is interesting, we may be able to do little about it since forgetting is part of human nature. The more interesting and policy relevant question is how much it costs to have high device variety conditional on knowing that surgeons do forget over time. To estimate the cost of forgetting on stems from this perspective, we first calculate the time gap between surgeries that use a stem\footnote{Some revision surgeries may not use a stem.} and the time gap between surgeries that use the same stem version in our data sample. On average, surgeons in our sample perform
a surgery using a stem once a week, and they use the same stem version once a month. Now consider a case where all stems belong to the same component version (no stem variety). By replacing the time gap between surgeries using the same stem version by the time gap between surgeries using a stem, we can calculate the hypothetical surgery duration in this case and compare it with the real surgery duration for each surgery.\footnote{Similar to the first case, we use the following two equations to calculate the hypothetical duration for each surgery using a stem:}

\[
\log (\text{duration} \_\text{real}) = x + 0.027 \times \log (\text{gap} \_\text{stem} + 1) \\
\log (\text{duration} \_\text{hypothetical}) = x + 0.027 \times \log (\text{gap} + 1)
\]

We find that in our sample period, about 1500 minutes in total are added to surgery duration due to the longer experience time gap associated with the high variety of stems, i.e. 643 minutes per year. Similarly, high liner variety adds 619 minutes per year to surgery duration. Alternatively, suppose the component-specific experience gap is reduced by half due to lower device variety (instead of no device variety). In this case, the time saved each year is 333 minutes for stems and 312 minutes for liners. Using similar estimates of per minute costs and total volumes as above, in the US such a reduction in stems and liner variety alone would result in about $13 million savings in costs excluding savings in surgeon and anaesthetist costs, or an opportunity loss of 1.1% more hip surgeries that could be performed, annually.

Given the extraordinarily high variety of component SKUs available and in use today for most medical devices, these findings have very significant implications for policy makers. The high productivity and quality costs associated with device variety suggest that the gain from a new device design needs to be large enough to compensate for the disadvantages of starting up on a new learning curve, and, also, of increasing the chances of knowledge deprecation over time.

In the current regulatory setup in the US, a device manufacturer can sidestep medical trials altogether and use a simple, six-month, 401K process to bring a new device to market, provided it can be shown to be "substantially equivalent" to a legally marketed existing device (Meier 2011). In the UK as well, 24% of all hip replacement devices available to surgeons have no evidence for their clinical effectiveness (Kynaston-Pearson et al. 2013). This regulatory environment also allows device manufacturers to quickly introduce very similar "me too" products. Further, individual surgeons have been known to work with device manufacturers to create minor new variants for which they provide sales avenues through their professional networks (Demske 2008). Our estimates highlight the costs of such extreme device proliferation.
An inherent problem in determining whether a new device has significantly better long-term outcomes is that one must observe patients over many years to answer this question. Early feedback on performance, typically measured by need for revision surgery, starts to flow in during the first few years of use of any new device. However, information on the field performance of new devices is often not readily available to surgeons. In recent years, several countries including Sweden, the UK, and Australia have started national joint registries, to record every usage of each device and to track performance and disseminate this information to the medical community. In the US, where there is no joint registry, information on the performance of new devices is hard to obtain. Our results suggest that timely sharing of outcome information can reduce both the short-term and long-term costs associated with unworthwhile variety.

Policy change can address the issue of better information about the effectiveness of new devices. At the same time, hospitals can reduce the costs of device variety through better surgical education. Many recent articles in the most prestigious medical journals have decried the current state of surgical education for its lack of structure and lack of innovation (e.g. Aggarwal and Darzi 2006). Our research highlights a specific need area – ways to adequately train surgeons on the wide variety of available device versions. In medical school, surgeons-in-training practice on cadavers and synthetic plastic bones using surgical tools and devices. Students are often taught using only one or two versions of a medical device, resulting in an a priori high probability that a surgeon will be using a device for the first time on a patient.\textsuperscript{34} Surges do gain more experience with the "see one, do one, teach one" apprenticeship model. But even for experienced surgeons such as those in our dataset, high device variety reduces the chances that a surgeon would have seen a particular device before.

Our orthopedic surgeon coauthor as well as other orthopedic surgeons we have spoken with have suggested some relatively easy ways to prepare for first use of a device version, such as carefully reading the documentation, examining the components themselves beforehand and talking to a colleague who has used the component before. Informal estimates based on our discussions with surgeons suggest that most surgeons do not take these preparatory steps.

Aside from such simple steps, new techniques such as surgical simulation, which can enable surgeons to practice on a wide variety of component versions prior to performing surgery, may reduce the costs of device variety. Prior research has shown that surgical residents trained using virtual reality simulation operate significantly faster and better than others trained using standard methods when doing the same standard procedure repeatedly using identical instruments (Saleh et al. 2009, Seymour et al. 2002). Seymour et al. (2002) mention the need for simula-

\textsuperscript{34}This information was shared with one of the authors by the Chief Information Officer of the Cleveland Clinic.
tions that train "very, very specific procedures rather than focus on acquisition of basic skills". Our quantitative estimates of the substantially longer duration associated with first time use of certain components highlight a key source of variation in duration even within a specific procedure, and underscore the need for on-the-job training whenever a surgeon encounters a new device version, for certain devices.

Our study has several limitations. We have focused on only one hospital and one type of surgery. Also, we have taken a reduced form approach in our empirical analysis. Future empirical work can use structural estimation to model surgeons’ choice of what device version to use in each surgery, that could formalize the selection and learning effects that we have found some evidence for. Future research can also consider behavioral aspects of component choice, in the spirit of emerging behavioral research in healthcare operations (e.g., Mennicken et al. 2014).

Our findings suggest that it is important to study learning and forgetting at the level of subprocesses, as knowledge gained and lost at this very granular level has significant implications for both cost and quality. While we have focused on healthcare, our modelling approach is relevant to any industry where products or services are comprised of numerous subprocesses. We also provide a general methodology to correct for the serious left censoring problem that commonly arises in this type of analysis.

References


**Appendix**

A.1: Estimate of Correlation Function \( \rho(d) \)

In this appendix, we show how to estimate a correlation function \( \rho(d_{t,v}) \) for two surgeries \( t \) and \( v \) by the same surgeon as a function of the time gap \( d_{t,v} = |t - v| \) (measured in days) between them.
Define $\tilde{e}_{st}$ as the residual from the pooled OLS presented by specification (3), $\tilde{e}_s$ as the sample mean of $\tilde{e}_{st}$ for each individual, and $\tilde{e}_{st}^*$ as the standardized residual which is

$$\tilde{e}_{st}^* = \frac{\tilde{e}_{st} - \tilde{e}_s}{\sigma^*},$$

where $\sigma^*$ is the standard deviation of $\tilde{e}_s$. A kernel-based estimate of the correlation function $\rho(d)$ is

$$\hat{\rho}(d) = \frac{\sum_s \sum_{t,v} K(d_{t,v} - d) \tilde{e}_{st} \tilde{e}_{sv}}{\sum_s \sum_{t,v} K(d_{t,v} - d)}$$ (11)

where $K(\cdot)$ is a kernel function and $d$ is its corresponding bandwidth. We use

$$K(z) = \begin{cases} \frac{h^{-1}}{\sqrt{2\pi}} \exp \left\{ -\frac{z^2}{2h^2} \right\} & \text{if } |z| \leq 4 \\ 0 & \text{if } |z| > 4 \end{cases}$$

and set $b = \sigma_d$. Note that the model which can be described by both equations (3) and (4) assumes a balanced panel, but the estimator for $\rho(d)$ in equation (11) does not require a balanced panel.

**A.2: Details of Simulation of $z^r_{st}$**

1. Estimate $\hat{F}_j(\cdot)$ for component $j$ by using Kaplan-Meier estimator;

2. $R$ random values $l^r$ are drawn from uniform distribution, and then used to find corresponding values $z^r = \hat{F}^{-1}_j(l^r)$;

3. For observations with $w_{sjt1} = 1$, we compare each $z^r$ with $w_{sjt2}$: if $z^r \leq w_{sjt2}$, then $(z^r_{sjt1} = 1, z^r_{sjt2} = 0)$; if $z^r > w_{sjt2}$, then $(z^r_{sjt1} = 1, z^r_{sjt2} = z^r)$. After the comparison, we have a matrix of $z^r_{sjt}$ with $2R$ elements;

4. For observations with $w_{sjt1} = 0$, we have a degenerate density $h_{j0}(z_{sjt1}, z_{sjt2} \mid w_{sjt1} = 0, w_{sjt2})$ and $(z^r_{sjt1} = 0, z^r_{sjt2} = w_{sjt1})$;

5. Do step 1-4 for each component to get $z^r_{st} = (z^r_{s1t}, z^r_{s2t}, z^r_{s3t}, z^r_{s4t})$ used in Equation (8).
A.3: Tables and Figures

Figure 1: Four Key Components Used in Hip Replacement Surgery

- Shell
- Liner
- Head
- Stem

Figure 2: Two Distinct Stem Component Versions

- Accolade Stem
- Secure-Fit Stem

Figure 3: Serial Correlation Function \( \rho(d) \)

\[
\begin{array}{c|c|c|c|c|c|c|c|c|c}
0 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & 0.3 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
20 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & 0.3 \\
40 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & 0.3 \\
60 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & 0.3 \\
80 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & 0.3 \\
100 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & 0.3 \\
\end{array}
\]
Figure 4: Data Censoring Problem for Surgeries with First Observed Use of a Component Variant

Figure 5: Kaplan-Meier Estimates of Distribution Functions for Experience Gap

Figure 6: The Effects of Learning and Forgetting (Stem)
Table 1: Comparison of Surgery Duration across Sub-Samples of Surgeries

<table>
<thead>
<tr>
<th></th>
<th>All Device Variants Observed in Use Before</th>
<th>At least 1 Device Variant Not Observed in Use Before</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Err.</td>
</tr>
<tr>
<td>Revision Surgeries</td>
<td>202.47</td>
<td>7.98</td>
</tr>
<tr>
<td>First-time Surgeries</td>
<td>146.89</td>
<td>2.84</td>
</tr>
</tbody>
</table>

Table 2: Information about Surgeons and Data Structure

<table>
<thead>
<tr>
<th>Surgeons</th>
<th>(1) Number of Surgeries in Sample Period</th>
<th>(2) Number of Surgeries Using at Least One Implant from 4 Main Vendors</th>
<th>(3) Estimation on Sample Size</th>
<th>(4) MD Completion Year</th>
<th>(5) Residency Completion Year</th>
<th>(6) Orthopedic Fellowship Completion Year</th>
<th>(7) Approximate Number of Hip Replacement Surgeries Pre-2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>365</td>
<td>350</td>
<td>268</td>
<td>1991</td>
<td>1996</td>
<td>1999</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>189</td>
<td>146</td>
<td>94</td>
<td>1991</td>
<td>1997</td>
<td>1999</td>
<td>1000</td>
</tr>
<tr>
<td>3</td>
<td>119</td>
<td>110</td>
<td>80</td>
<td>1993</td>
<td>1999</td>
<td>2000</td>
<td>500</td>
</tr>
<tr>
<td>4</td>
<td>79</td>
<td>65</td>
<td>41</td>
<td>1984</td>
<td>2005</td>
<td>2006</td>
<td>_</td>
</tr>
<tr>
<td>Total</td>
<td>752</td>
<td>671</td>
<td>483</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Table 3: Groupings of the Four Main Components Used in Hip Replacement Surgeries

<table>
<thead>
<tr>
<th>Company</th>
<th># of SKUs</th>
<th># of Component Variants</th>
<th># of SKUs</th>
<th># of Component Variants</th>
<th># of SKUs</th>
<th># of Component Variants</th>
<th># of SKUs</th>
<th># of Component Variants</th>
<th># of SKUs</th>
<th># of Component Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zimmer</td>
<td>16</td>
<td>4</td>
<td>10</td>
<td>9</td>
<td>19</td>
<td>6</td>
<td>20</td>
<td>4</td>
<td>65</td>
<td>23</td>
</tr>
<tr>
<td>Depuy</td>
<td>45</td>
<td>8</td>
<td>60</td>
<td>10</td>
<td>61</td>
<td>10</td>
<td>63</td>
<td>11</td>
<td>229</td>
<td>39</td>
</tr>
<tr>
<td>Stryker</td>
<td>31</td>
<td>5</td>
<td>48</td>
<td>12</td>
<td>32</td>
<td>11</td>
<td>48</td>
<td>15</td>
<td>159</td>
<td>43</td>
</tr>
<tr>
<td>Smith</td>
<td>22</td>
<td>3</td>
<td>44</td>
<td>4</td>
<td>10</td>
<td>1</td>
<td>34</td>
<td>8</td>
<td>110</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>114</td>
<td>20</td>
<td>162</td>
<td>35</td>
<td>122</td>
<td>28</td>
<td>165</td>
<td>38</td>
<td>563</td>
<td>121</td>
</tr>
</tbody>
</table>

Note: based on the 671 data sample used to create component version variables

Table 4: Covariance Matrix of Experience Variables

<table>
<thead>
<tr>
<th></th>
<th>Total Experience</th>
<th>First Time Use Dummy</th>
<th>Lin(Experience Gap)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Experience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shell</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stem</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liner</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Time Use Dummy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shell</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stem</td>
<td>-0.15</td>
<td>0.26</td>
<td>1</td>
</tr>
<tr>
<td>Liner</td>
<td>-0.11</td>
<td>0.22</td>
<td>0.09</td>
</tr>
<tr>
<td>Head</td>
<td>-0.17</td>
<td>0.26</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>-0.22</td>
<td>-0.22</td>
<td>-0.22</td>
</tr>
<tr>
<td></td>
<td>-0.04</td>
<td>-0.05</td>
<td>-0.31</td>
</tr>
<tr>
<td></td>
<td>0.11</td>
<td>-0.04</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td>-0.04</td>
<td>-0.09</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>-0.33</td>
<td>-0.04</td>
<td>-0.11</td>
</tr>
</tbody>
</table>

0.28 0.19 0.11 1
Table 5: Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration (minutes)</td>
<td>164.98</td>
<td>70.47</td>
<td>47</td>
<td>557</td>
</tr>
<tr>
<td>Total Experience</td>
<td>141.97</td>
<td>104.56</td>
<td>0</td>
<td>364</td>
</tr>
<tr>
<td>Variable: First Time Use Dummy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shell</td>
<td>0.05</td>
<td>0.21</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Stem</td>
<td>0.09</td>
<td>0.29</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Liner</td>
<td>0.08</td>
<td>0.28</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Head</td>
<td>0.10</td>
<td>0.30</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Variable: Experience Gap(days)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shell</td>
<td>24.44</td>
<td>47.87</td>
<td>0</td>
<td>395</td>
</tr>
<tr>
<td>Stem</td>
<td>30.08</td>
<td>64.57</td>
<td>0</td>
<td>521</td>
</tr>
<tr>
<td>Liner</td>
<td>26.02</td>
<td>52.08</td>
<td>0</td>
<td>378</td>
</tr>
<tr>
<td>Head</td>
<td>42.46</td>
<td>85.45</td>
<td>0</td>
<td>636</td>
</tr>
<tr>
<td>Patient Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.49</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BMI</td>
<td>29.90</td>
<td>7.02</td>
<td>15.44</td>
<td>62.13</td>
</tr>
<tr>
<td>Age</td>
<td>60.34</td>
<td>13.65</td>
<td>22</td>
<td>91</td>
</tr>
<tr>
<td>ASA Average</td>
<td>2.43</td>
<td>0.51</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td># of Comorbidities</td>
<td>1.99</td>
<td>1.45</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Surgery Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Legs</td>
<td>0.01</td>
<td>0.09</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Unileg</td>
<td>0.07</td>
<td>0.25</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cemented</td>
<td>0.16</td>
<td>0.37</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Use: shell</td>
<td>0.86</td>
<td>0.35</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Use: stem</td>
<td>0.84</td>
<td>0.36</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Use: liner</td>
<td>0.72</td>
<td>0.45</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Use: head</td>
<td>0.97</td>
<td>0.17</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reason: Revision</td>
<td>0.24</td>
<td>0.43</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reason: AVN</td>
<td>0.11</td>
<td>0.31</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reason: Displasia</td>
<td>0.04</td>
<td>0.20</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reason: Arthritis</td>
<td>0.58</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reason: Severe Arthritis</td>
<td>0.04</td>
<td>0.19</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reason: End Stage Arthritis</td>
<td>0.05</td>
<td>0.23</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reason: Fracture</td>
<td>0.07</td>
<td>0.26</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reason: Other</td>
<td>0.05</td>
<td>0.21</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Reasons for revision</td>
<td>0.17</td>
<td>0.40</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Notes: # of surgeons = 4; # of observations = 483
### Table 6: Estimation Results - Dependent Variable is Ln(Duration)

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>Column (1)</th>
<th>Column (2)</th>
<th>Column (3)</th>
<th>Column (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS: Total Experience</td>
<td>OLS: also including surgeon FE</td>
<td>OLS: also including Component Experience</td>
<td>MLE with Simulation</td>
</tr>
<tr>
<td></td>
<td>Coef. (Std. Err.)</td>
<td>Coef. (Std. Err.)</td>
<td>Coef. (Std. Err.)</td>
<td>Coef. (Std. Err.)</td>
</tr>
<tr>
<td>Total Experience/100</td>
<td>-0.110*** (0.025)</td>
<td>0.026 (0.050)</td>
<td>0.034 (0.050)</td>
<td>0.032 (0.062)</td>
</tr>
<tr>
<td>First Time Use Dummy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shell</td>
<td>0.136 (0.092)</td>
<td>0.161* (0.098)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stem</td>
<td>0.238*** (0.072)</td>
<td>0.259*** (0.061)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liner</td>
<td>0.068 (0.077)</td>
<td>0.066 (0.079)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>-0.038 (0.068)</td>
<td>-0.036 (0.075)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log(Experience Gap)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shell</td>
<td>-0.009 (0.015)</td>
<td>-0.895 (1.679)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stem</td>
<td>0.027** (0.013)</td>
<td>0.027** (1.343)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liner</td>
<td>0.028* (0.015)</td>
<td>0.027* (0.016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>0.003 (0.012)</td>
<td>0.265 (1.439)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.101*** (0.030)</td>
<td>0.099*** (0.030)</td>
<td>0.107*** (0.030)</td>
<td>0.106*** (0.032)</td>
</tr>
<tr>
<td>BMI/100</td>
<td>0.299 (0.219)</td>
<td>0.313 (0.216)</td>
<td>0.421*** (0.216)</td>
<td>0.425*** (0.186)</td>
</tr>
<tr>
<td>Age/100</td>
<td>-0.522*** (0.128)</td>
<td>-0.472*** (0.127)</td>
<td>-0.428*** (0.126)</td>
<td>-0.420*** (0.117)</td>
</tr>
<tr>
<td>ASA Average</td>
<td>0.006 (0.034)</td>
<td>0.012 (0.034)</td>
<td>0.002 (0.033)</td>
<td>0.002 (0.036)</td>
</tr>
<tr>
<td># of Comorbidities</td>
<td>0.041*** (0.012)</td>
<td>0.036*** (0.012)</td>
<td>0.033*** (0.012)</td>
<td>0.033*** (0.013)</td>
</tr>
<tr>
<td>Surgery Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Legs</td>
<td>0.681*** (0.160)</td>
<td>0.700*** (0.158)</td>
<td>0.715*** (0.157)</td>
<td>0.719 (0.579)</td>
</tr>
<tr>
<td>Unihead</td>
<td>-0.113 (0.070)</td>
<td>-0.069 (0.070)</td>
<td>-0.083 (0.070)</td>
<td>-0.087 (0.074)</td>
</tr>
<tr>
<td>Cemented</td>
<td>0.053 (0.043)</td>
<td>0.004 (0.048)</td>
<td>-0.034 (0.049)</td>
<td>-0.034 (0.052)</td>
</tr>
<tr>
<td>Reason: Revision</td>
<td>0.345*** (0.075)</td>
<td>0.370*** (0.075)</td>
<td>0.302*** (0.077)</td>
<td>0.302*** (0.067)</td>
</tr>
<tr>
<td>Reason: Avascular Necrosis</td>
<td>-0.102 (0.068)</td>
<td>-0.096 (0.067)</td>
<td>-0.066 (0.066)</td>
<td>-0.065 (0.079)</td>
</tr>
<tr>
<td>Reason: Displasia</td>
<td>0.093 (0.074)</td>
<td>0.099 (0.073)</td>
<td>0.070 (0.074)</td>
<td>0.070 (0.094)</td>
</tr>
<tr>
<td>Reason: Arthritis</td>
<td>-0.062 (0.072)</td>
<td>-0.058 (0.072)</td>
<td>-0.027 (0.072)</td>
<td>-0.025 (0.074)</td>
</tr>
<tr>
<td>Reason: Severe Arthritis</td>
<td>0.072 (0.098)</td>
<td>0.092 (0.099)</td>
<td>0.133 (0.098)</td>
<td>0.130 (0.113)</td>
</tr>
<tr>
<td>Reason: End Stage Arthritis</td>
<td>0.071 (0.094)</td>
<td>0.053 (0.093)</td>
<td>0.053 (0.092)</td>
<td>0.054 (0.114)</td>
</tr>
<tr>
<td>Reason: Fracture</td>
<td>-0.008 (0.078)</td>
<td>-0.015 (0.077)</td>
<td>-0.032 (0.078)</td>
<td>-0.036 (0.077)</td>
</tr>
<tr>
<td>Reason: Other</td>
<td>-0.023 (0.106)</td>
<td>-0.061 (0.105)</td>
<td>-0.014 (0.106)</td>
<td>-0.017 (0.097)</td>
</tr>
<tr>
<td>Reasons for revision</td>
<td>-0.069 (0.082)</td>
<td>-0.059 (0.081)</td>
<td>-0.020 (0.082)</td>
<td>-0.020 (0.078)</td>
</tr>
<tr>
<td>Time Trend</td>
<td>1.004*** (0.261)</td>
<td>0.621** (0.293)</td>
<td>0.710** (0.297)</td>
<td>0.682* (0.366)</td>
</tr>
<tr>
<td>Quadratic Time Trend</td>
<td>-0.127 (0.277)</td>
<td>-0.197 (0.276)</td>
<td>-0.381 (0.279)</td>
<td>-0.294 (0.319)</td>
</tr>
</tbody>
</table>

- # of Observations: 483
- Adj. R-squared: 0.381, 0.397, 0.417

**Note:** All regressions include a dummy for use of each component type (and a dummy for each reason for surgery). The time trend is defined as the number of days since start of the sample period divided by 1000.

*** p<0.01, ** p<0.05, * p<0.1