Three Essays on Learning and Collaboration in Operations Management

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Operations Management

By

Yue Yin

EVANSTON, ILLINOIS

September 2019
© Copyright by Yue Yin 2019
All Rights Reserved
ABSTRACT

Three Essays on Learning and Collaboration in Operations Management

Yue Yin

In knowledge-intensive work such as data analysis or emergency medicine, learning and collaboration is integral to the effective “production” process. Uncovering the mechanisms, benefits and costs of learning and collaboration is crucial for the design and operation of a successful process. My Ph.D. research centers on empirically investigating the ways in which learning and collaboration among people affect the performance of the processes they execute. In this collection of essays, I study some of the effects of three modes of learning and collaboration (learning by doing, learning from peers and supervision) on operational performance, together with my advisors at Kellogg and collaborators at eBay and NorthShore Health System. This dissertation incorporates two published papers: Chapter 1 uses material from [Yin et al. (2018)] while Chapter 3 uses material from [Wang et al.].

In Chapter 1, we investigate how data-analyst productivity benefits from collaborative platforms that facilitate learning-by-doing (i.e. analysts learning by writing queries
on their own) and learning-by-viewing (i.e. analysts learning by viewing queries written by peers). Learning is measured using a behavioral (productivity-improvement) approach. Productivity is measured using the time from creating an empty query to first executing it. Using a sample of 2,001 data analysts at eBay Inc. who have written 79,797 queries from 2014 to 2018, we find that: 1) learning-by-doing is associated with significant productivity improvement when the analyst’s prior experience focuses on the focally queried database; 2) only viewing queries that are authored by analysts with high output rate (average number of queries written per month) is associated with significant improvement in the viewer’s productivity; 3) learning-by-viewing also depends on the “social influence” of the author of the viewed query, which we measure ‘locally’ based on the number of the author’s direct viewers per month or ‘globally’ based on the how the author’s queries propagate to her peers in the overall collaboration network. Combining results 2 and 3, when segmenting analysts based on output rate and ‘local’ social influence, the viewing of queries authored by analysts with high output but low local influence is associated with the largest improvement in the viewer’s productivity; whereas when segmenting based on output rate and ‘global’ social influence, the viewing of queries authored analysts with high output and high global influence is associated with the largest improvement in the viewer’s productivity. Overall, regardless of the segmentation, learning-by-viewing is associated with greater productivity improvement than learning-by-doing in our study.

In Chapter 2, we investigate whether an individual’s productivity and learning ability vary over time. A Hidden Markov Model (HMM) is proposed to capture such dynamics, and is applied to the same empirical setting as in chapter 1. This model
enables us to segment data analysts into several latent states with respect to their productivity and learning ability. These analysts are allowed to transit between states, the direction and probability of which depend on their participation in two modes of learning activities (writing own queries and viewing peers’ queries). The effect of an analyst’s participation in either kind of learning activity also varies by her state. We find that the three-state HMM model fits better than the standard learning curve model in relevant measures. This implies that the dynamic model is more appropriate to capture the evolvement of analysts’ productivity and learning. We have identified three latent states (novice, intermediate and advanced) in ascending order of the intrinsic productivity of analysts. Our findings reveal different patterns of learning in different states. Only analysts in the novice state benefit from both writing own queries and viewing peers’ queries. The learning effects from writing own queries decreases with higher state. We also find that analysts in intermediate state or above risk transiting to lower states by viewing peers’ queries.

In Chapter 3, we perform an observational, time-motion study on 25 EPs who worked in a community-academic ED and a non-academic community ED. The objective is to compare attending emergency physician (EP) time spent on direct and indirect patient care activities in emergency departments (ED) with, and without, emergency medicine (EM) residents. Two observations of each EP were performed at each site. Average time spent per 240-minute observation on main-category activities are summarized in percentages. We report descriptive statistics (median and interquartile ranges) for the number of minutes EPs spent per sub-category activity, in total and per patient. We performed a Wilcoxon two-sample test to assess differences between time
spent across two EDs. The results show that the 25 observed EPs executed 34,358 tasks in the two EDs. At the community-academic ED, EPs spent 14.2% of their time (8.5 minutes/hour) supervising EM residents. Supervision activities included data presentation, medical decision making, and treatment. The time spent on supervision is offset by a decrease in time spent by EPs on indirect patient care at the community-academic ED compared to the non-academic community ED, specifically from communication and EHR work. There was no statistical difference with respect to direct patient-care time expenditure across two EDs. There was a non-statistically significant difference in attending patient load between sites.
Acknowledgements

Special thanks is reserved for my Ph.D. advisors, Professor Jan A. Van Mieghem and Professor Itai Gurvich. I highly appreciate all their contributions of time, ideas and funding to complete my Ph.D. experience and to allow me grow as an independent researcher. I also owe a sincere thank you to Professor Martin Lariviere and Professor Robert Bray, for serving as my committee members and letting my Ph.D. defense be an enjoyable moment. Thank you, for all the care and support as well as endless brilliant comments and suggestions. I am also very grateful to Professor Dennis Zhang for his scientific advice, knowledge, and all the effort trying to put me on the right track.

I would like to express my gratitude and appreciation for Dr. Ernest E. Wang for initiating and coordinating the research project in emergency department. The meetings and conversations were vital in educating me to on healthcare research. I would also like to extend thanks to the physicians, residents and staffs in Glenbrook and Highland Park hospitals at NorthShore Health System. All of you have been there to support me when I collected data for my research project.

I am deeply grateful to Debora Seys for all her continuous trust, help and encouragement throughout the research project in collaboration with eBay. I think of her as a loving and caring aunt who always has positive influence on my life.

Many thanks to my friends together with me on the journey of Ph.D.: Subing Qu, Haoxiang Yang, Nil Karacaoglu. I have been lucky to have a chance to complete this
journey with your companion. I will forever remember all the coffee, drinks and midnight talk.

I owe a huge debt of gratitude to my parents for immense support, love, and patience. Without them, this dissertation would never have been written. Thanks a lot to my significant other Ruozhou Yan, for sitting next to me for hours, and for providing guidance, delicious meals and great music when required.
Table of Contents

**ABSTRACT** 3

Acknowledgements 7

Table of Contents 9

List of Tables 11

List of Figures 13

Chapter 1. Learning by Doing versus Learning by Viewing: An Empirical Study of Data Analyst Productivity on a Collaborative Platform at eBay

(Joint with Jan A. Van Mieghem and Itai Gurvich) 14

1.1. Introduction 14

1.2. Theory and Hypotheses 19

1.3. Methods 27

1.4. Analysis and Results 36

1.5. Summary, Discussion, and Limitations 48

1.6. Conclusion 56

Chapter 2. A Hidden Markov Model of Data Analyst Productivity and Learning Dynamics

(Joint with Jan A. Van Mieghem and Itai Gurvich) 57
List of Tables

1.1 Descriptive statistics of the raw data capturing \( N = 79797 \) queries written and executed by 2001 data analysts at eBay during 2014-2018. 37

1.2 Results of Zero-truncated Negative Binomial Regressions for Model 1-2 40

1.3 Results of Zero-truncated Negative Binomial Regressions for Model 3-6 41

1.4 Comparison of Zero-truncated Negative Binomial Regression Models 42

1.5 ANOVA results reject both \( \beta_1 = \beta_2 \) and \( \beta_3 = \beta_4 = \beta_5 = \beta_6 \) 44

1.6 Variance Inflation Factor of Explanatory Variables 45

2.1 Fit Measures for Comparing Models 79

2.2 Estimation Results for the Three-state Hidden Markov Models Parameter 80

3.1 Performance Metrics for community-academic and community in FY2017 89

3.2 Performance Metrics for community-academic and community in FY2017 92
3.3 Shapiro-Wilk Test for Normality of Total Minutes Spent on Sub-category at community-academic vs. community

3.4 Shapiro-Wilk Test for Normality of Average Minutes Spent per Patient on Sub-category at community-academic vs. community

3.5 Average Minutes per Patient (during the 240-minute Observation) Spent on Sub-category at community-academic vs. community
## List of Figures

1.1 Segmenting data analysts using two dimensions—output rate and social influence—into four types: All-star, Lone-wolf, Maven and Non-star. 25

1.2 Similar to query writing, a software development process typically starts with a "Pre-alpha" stage. 30

1.3 The First Completion Time is the first stage of a typical query process. 31

1.4 Segmenting data analysts by Output-rate × Viewership 34

1.5 Segmenting data analysts by Output-rate × PageRank 35

2.1 A Hidden Markov Model of Data Analyst Learning and Writing Query 65

3.1 Mean Task Percentages for the 240-minute observations 98
CHAPTER 1

Learning by Doing versus Learning by Viewing: An Empirical Study of Data Analyst Productivity on a Collaborative Platform at eBay

(Joint with Jan A. Van Mieghem and Itai Gurvich)

1.1. Introduction

Effective data analytics drives business success by enhancing managerial decision-making. Companies often struggle to maintain growth in the productivity of their data analysts. In this paper, we investigate how data-analyst productivity benefits from collaborative platforms that, in addition to providing a query-writing environment for the analysts, also facilitate access to queries authored by their peers. Productivity is measured using the time from creating an empty query to first executing it.

Companies in various industries employ data analysts. From financial services to retailing to social media, data analysts aim to transform data into valuable information for decision-makers. By retrieving, organizing and narrating raw data, data analysts can spot trends in the market, enable managers to know more about their consumers, and develop recommendations that are aligned with the company’s strategy.

The current demand for skilled data analysts outpaces the supply. A 2016 McKinsey
Global Institute report concludes that by 2024 the U.S. economy would be in shortage of 250,000 analytics professionals (Henke et al. 2017). Furthermore, surveyed business leaders find it challenging to recruit and retain proficient data analysts (Harrington 2017).

A good data analyst must be technologically up-to-date to bring insightful results swiftly (Chen et al. 2012). The writing and execution of queries is central to the work of data analysts with large-scale data. Queries are written using SQL (Structured Query Language) (Pavlo et al. 2009) to answer business questions like: What goods are flying off a retailer’s shelf? Who prefers to shop in a boutique store versus on-line? What are the ten most frequently searched items on eBay Motors in London in July 2017? Data analysts answer these questions by extracting data from the proper databases, performing data manipulations (e.g., sorting, grouping, filtering and joining), and finally reporting the results.

Proficiency in programming SQL queries is a learned skill. Organizational learning theory defines learning as a change in the knowledge that occurs as a function of experience (Fiol and Lyles 1985). Such knowledge transformation occurs at different levels in organizations — individual, group, organizational, and inter-organizational (Crossan et al. 1999). Studies have demonstrated that much of the programming knowledge is tacit knowledge, i.e., knowledge that usually is not openly expressed or taught (Soloway et al. 1982; Wagner and Sternberg 1985). It also has been established that expert programmers know more than mere syntax and semantics of a particular language; compared to novices, their knowledge is better organized. For example, an
expert programmer would be able to see the underlying commonalities and the differences among various problems and programs. Numerous researchers who aim to characterize expert programming also have suggested the existence of reusable “chunks” of knowledge representing solution patterns that achieve different kinds of goals (Lane and VanLehn 2005).

Most current approaches measure learning by assessing changes in cognition. Qualitative methods like questionnaires, interviews and verbal-protocol analyses are typically used to that end (Ericsson and Simon 1980; Huff and Jenkins 2002). Such cognitive approaches are nonetheless unable to capture tacit knowledge (Hodgkinson and Sparrow 2002). Behavioral approaches measure learning by assessing changes in practices or performance and have been shown to capture the tacit knowledge well (Argote 2012; Argote and Epple 1990; Dutton and Thomas 1984). Recently more researchers start exploring a behavioral approach to quantitatively measure such ‘informal learning’ from a large online community. For example, Yang et al. measured learning of a Scratch user as growth in the cumulative repertoire of weighted vocabulary block use (Yang et al. 2015; Dasgupta et al. 2016). We thereby deploy a behavioral approach—measuring change in performance—in our study. The expedition of SQL programming indicates the positive learning outcome of data analysts.

In organizational learning theory, experience—typically defined as the total or cumulative number of task completions—underpins learning. The most fundamental characterization of experience is whether it is acquired directly by the focal organizational unit or indirectly from other units (Argote 2012). Two modes of organizational learning are derived from this characterization: learning from direct experience (i.e.
from one’s own practices) and learning from indirect experience (i.e. from other organizational units’ experience) (Kozlowski 2012). Learning from direct experience is the embodiment of “practice makes perfect” while learning from indirect experience emphasizes the circulation of knowledge among peers. Studies of the well-known learning curve provide considerable evidence of learning from direct experience (Yelle 1979; Dutton and Thomas 1984). There is also extensive work on collaboration and knowledge sharing that investigates learning from indirect experience (Levitt and March 1988; Huber 1991; Miner and Haunschild 1995; Gino et al. 2010; Argote et al. 2000).

We study these two modes of organizational learning when eBay data analysts work on Alation. Alation is an enterprise collaborative data platform that makes data accessible to individuals across the organization. The platform empowers analysts to write SQL queries using well-curated data, and allows them to publish their own queries or view any public query authored by their peers. We focus on two potential learning processes on Alation: learning-by-doing (“by oneself”) vs. learning-by-viewing (peers’ queries). Learning-by-doing captures how data analysts become faster the more queries they write; how they learn from direct experience. In contrast, learning-by-viewing captures how data analysts improve by viewing queries authored by their peers; it is an instance of learning from indirect experience. We pose the following research questions:

1. Learning-by-doing: How is data analyst productivity associated with self-practice?
2. Learning-by-viewing: How is data analyst productivity associated with viewing of queries authored by peers?
Organizational learning theory emphasizes the role of expert in facilitating the diffusion and validation of credible knowledge. Previous studies have demonstrated that group members are likely to accept and put more weight on information from a recognized expert (Stasser et al. 1995). The retaining of and the interaction with exceptional performers appears to affect organizational outcomes (Argote 2012; Burt 2009). Given that most of the programming knowledge is tacit knowledge, it is not clear how to characterize expert (or “star”) analysts. The conventional approaches to the characterization of stars are based exclusively on individual output (Zucker et al. 1998; Groysberg et al. 2008; Azoulay et al. 2010). Classic economic growth theories nonetheless claim that human capital externalities (e.g. the influence that an individual has on the performance of others) are also a key input in the generation of knowledge (Romer 1990; Acemoglu 1996; Lucas Jr 1988). Recent empirical work adopts this view and expands the traditional characterization of ‘star’ by adding measurements of social influence (Oettl 2012). Recent studies on software development gauge expertise identification approaches by considering the code a developer authors and the code that the developer consults during their work (Fritz et al. 2010; Schuler and Zimmermann 2008).

Inspired by the literature on the role of experts in organizational learning, we ask:

3. **Learning from experts**: How is learning-by-viewing a query associated with the expertise of the query’s author?

Our study is also related to two well-known perspectives in cognitive psychology and learning sciences (Anderson et al. 1997; Greeno 1997; Cobb and Bowers 1999): the “within-the-human” perspective that is generally attributed to Jean Piaget (Piaget 1975) vs. the situated cognition perspective proposed by Jean Lave and Etienne Wenger.
Lave et al. [1991]). The former focuses on the internalized development within the individual’s mental representation of the world, and presumes that knowledge can be constructed through one’s own practices (Cobb and Bowers 1999). The latter, in contrast, suggests that knowledge is constructed in a social context, and individual participation in valued social practices is critical for successful learning (Cobb and Bowers 1999; Greeno 1997). Read in our context, the within-the-human perspective would suggest that, mostly, the analyst acquires programming knowledge from writing queries by herself (learning-by-doing). The situated-cognition perspective suggests, instead, that learning is mostly accomplished through the analyst’s interaction with other analysts (learning-by-viewing). For situated-cognition perspective it is needed, of course, for newcomers to be able to observe experts. In our context Alation is the platform that facilitates this so-called “legitimate peripheral participation” (Lave et al. 1991).

The rest of the paper is organized as follows. In the next section, we discuss the theory and develop our hypotheses. We then describe the empirical setting, data and measures. Later, we present our analysis strategy and report our results. We conclude with a comprehensive discussion of the results and their potential implications.

1.2. Theory and Hypotheses

Organization learning theory has been applied in a broad spectrum of industries from manufacturing to services (Benkard 1999; Darr et al. 1995). Other recent studies focus on knowledge industries like IT consulting and software development (Fong Boh et al. 2007; Kim et al. 2012; Kang and Hahn 2009). We follow this path, relying on organizational learning theory to develop and test hypotheses related to the learning
of data-analysts working, as individuals, on a collaborative platform that facilitates knowledge sharing. Our hypotheses pertain to two modes of organizational learning: learning from direct experience and learning from indirect experience (Reagans et al. 2005; Fong Boh et al. 2007; Kc et al. 2013). We use productivity as the main variable of interest and measure how it relates to the accumulated experience that data analysts have gained by writing queries on their own (learning-by-doing) and by viewing queries written by their peers (learning-by-viewing). In formulating the hypotheses we also rely on two cognitive theories of learning: learning-by-doing as it relates to the within-the-human perspective and learning-by-viewing as it relates to situated-learning perspective.

1.2.1. Learning by Doing

1.2.1.1. Individual Learning from Direct Experience. Recent studies of individual learning from direct experience cover various industries. Kim et al. estimate the learning curve of IT consultants (Kim et al. 2012).KC et al. examine the direct impact of a cardiologist’s own prior experience on individual learning (Kc et al. 2013). Staats and Gino compare the benefits of individual worker’s experience in a day or over several days (Staats and Gino 2012). They find that specialization is related to productivity improvement over a single day. We expect data analysts in our study to benefit from their past experience of writing queries. A positive answer to the following hypothesis further validates the earlier findings and extends them to the context of data analysts.

Hypothesis 1. Past experience of writing queries is associated with an improvement in the data analyst’s productivity.
1.2.1.2. Specificity of Direct Experience. Repetition of a given task is likely to improve the performance of an individual more than experience with related (but different) tasks. Boh et al. show that specialized experience with the same system has the greatest impact on productivity for modification requests completed by individual developers (Kang and Hahn 2009; Fong Boh et al. 2007). Such findings, identifying the benefit of focal experience, appear in other service industries (Kc and Staats 2012; Staats and Gino 2012; Egelman et al. 2016).

On Alation, data analysts write queries using different databases. In an interview study by Kandel et al. focusing on the challenges of data analysts, most of the respondents mentioned the difficulty in interpreting certain database fields (Kandel et al. 2012). The evidence in the literature suggests that, as the analyst practices more with the focal database, she will become more familiar with this database’s nuances, including the field definitions, data quality and assumptions. We therefore measure the association between productivity and specificity of experience. A positive answer to the hypothesis below corroborates the value of focal direct experience in writing queries.

Hypothesis 2. Past experience in querying the focal database is associated with greater improvement in data analyst productivity than past experience querying different databases.

1.2.2. Learning by Viewing

1.2.2.1. Individual Learning from Indirect Experience. "social interaction among individuals, groups and organizations are fundamental to organizational knowledge creation” (Nonaka 1994). Organizational learning is frequently an interactive, social phenomenon (Tyre and Von Hippel 1997). Such communal processes are important
because no one person embodies sufficient knowledge for solving all complex organizational problems. For instance, in the context of machine repair technicians, most of the knowledge is not acquired in the classroom, but comes, rather, from informal story-sharing among technicians and users about their experiences in particular work environments (Orr, 2016; Brown and Duguid, 2000). This finding, that individuals also benefit from their peers’ experience (learning from indirect experience) is also confirmed in (Levitt and March, 1988; Huber, 1991; Miner and Haunschild, 1995; Gino et al., 2010; Huckman and Staats, 2011; Narayanan et al., 2014; Hwang et al., 2015; Reagans et al., 2016; Staats, 2012).

In the context of computer programming, Brandt et al. (Brandt et al., 2009) propose that by relying on information and source code fragments provided by other people from the Web, developers engage in just-in-time learning of new skills and approaches, clarify and extend their existing knowledge, and remind themselves of details deemed not worth remembering. Vasilescu et al. (Vasilescu et al., 2013) argue that participation in on-line programming communities (e.g. StackOverflow) speeds up code development since quick solutions to technical challenges can be provided by peers. Dasgupta et al. confirmed that remixing—defined as the reworking and combination of existing creative artifacts—acts a pathway to learning (Dasgupta et al., 2016). They found that a learner’s repertoire of programming concepts increases when she engages in remixing.
Yet there is a trade-off. Viewing peers’ code may delay programming activities as both viewing and programming compete for the developer’s time and attention. Current empirical evidence for the benefit of learning from indirect experience is inconclusive. Waldinger finds no evidence for peer effects on the productivity of researchers in physics, chemistry and mathematics. In his study, even very high-quality scientists do not affect the productivity of their local peers (Waldinger 2011). KC et al. investigate the relationship between cardiologists’ current performance and the performance of their colleagues in the same hospital (Kc et al. 2013). But their data does not include detailed “views” information, i.e., what information individual cardiologists actually observe or share among each other. The authors, therefore, call for future research to identify the precise micro-mechanisms at work, exploring how knowledge is shared among individuals and affects their performance. Our fine-grained data include a complete history of each analyst’s record of viewing specific peers’ queries, offering an opportunity to respond to the authors’ call. We test the following hypothesis to measure learning from indirect experience in analysts writing queries. A positive answer to this hypothesis confirms that it is highly possible that analysts who mostly view queries written by peers bear high productivity. A negative answer still leaves the possibility that only viewing queries written by certain peers predicts high productivity. This is investigated in section 2.2.2.

**Hypothesis 3.** Past experience of viewing queries written by peers is positively associated with data analyst productivity.
1.2.2.2. Characterizing Star Data Analysts. Previous studies on knowledge spillover and peer effects among scientists suggest a differential impact of collaborating with different types of individuals. Exceptional performers, or stars, may greatly advance the production of ideas and the innovation process (Azoulay et al. 2010; Oettl 2012). The situated-cognition theory in learning sciences places and emphasizes on the role of experts (Dede et al. 2004). In the context of programming and software development, developers appear to have greater interest in following some prolific developers, who are considered ‘coding rockstars’ by the overall community (Dabbish et al. 2012). In the on-line programming community, a developer’s status can affect decision-making. Tsay et al. find that contributions from higher-status submitters are more readily accepted by project managers (Tsay et al. 2014).

In the context of data analysts, we must first ask how to identify (or characterize) the “rockstars” and, given such a characterization, how are these said stars associated with the changes in productivity of their peers? The characterization we propose in this paper differs from most existing taxonomy by considering not just the individual output (Zucker et al. 1998; Groysberg et al. 2008; Azoulay et al. 2010), but also the individual’s social influence on her peers. Such an expanded definition is unavoidable here as we wish to measure learning-by-viewing which is an interaction-based construct. This need is also identified in (Tsay et al. 2014; Oettl 2012) who suggest that the characterization of “stars” should consider the individuals’ social influence (Tsay et al. 2014; Oettl 2012). We introduce a two-dimensional segmentation of analysts that incorporates both a measure of the individual analyst’ output and a measure of her social influence on Alation. Relying on Oettl’s characterization of star scientists (Oettl 2012) we
Figure 1.1. Segmenting data analysts using two dimensions—output rate and social influence—into four types: All-star, Lone-wolf, Maven and Non-star.

We specify two segmentations in this paper that both use output rate yet each use a different measure of social influence. We adopt the conventional measure of individual output: \( output-rate = \) the average number of queries created by the analyst \( i \) per unit of time. \cite{Adams2009,Groysberg2008} Taking months as unit of time, we will use:

\[
(1.1) \quad \text{Monthly Output of Queries}_i = \frac{\text{Total number of queries } i \text{ has written}}{\text{Number of months since } i \text{ joined in } Alation}
\]
We adopt two different measures of social influence, *viewership* and *PageRank*, to describe how influential an analyst’s queries have been since they were written on *Alation*. *Viewership* captures the average number of distinct viewers per month of all queries authored by analyst $i$. We define $N_{\text{distinct viewers},i}$ as the total number of distinct data analysts who have viewed $i$’s queries and $T_{\text{livetime},k}$ as the number of months that query $k$ is viewable:

\[
\text{Monthly Viewers per Query}_i = \frac{N_{\text{distinct viewers},i}}{\sum_{\text{query } k \text{ written by } i} T_{\text{livetime},k}}
\]

which represents the attention that focal analyst $i$ receives from her peers on *Alation*. Viewership is a measure of the "local" influence of an author on its direct viewers. A qualitative interview study by Dabbish et al. (Dabbish et al. 2012) demonstrates that, in large-scale distributed collaborations and communities of practice (e.g. GitHub), the attention that a developer has received signals her status in the community. Quoting to a representative participant in their study, “[One visible cue is] the number of people watching a project or people interested in the project; obviously it’s a better project than versus something that has no one else interested in it.”

The second measure of social influence is *PageRank*, which was proposed by Google for weighting the importance of a web page based on the number and quality of links to this page (Brin and Page 1998). To compute the *PageRank* of each data analyst in our study, we first build a directed, analyst-to-analyst network that represents the social interactions on *Alation*, which we explain later in 3.2.2. Then, running the PageRank algorithm on this network returns the *PageRank* for every data analyst. The analyst with higher *PageRank* is considered more influential on the overall network herself.
While viewership captures the ‘local’ influence, PageRank represents a ‘global’ network-wide influence of an analyst by capturing not only her direct viewers but also the viewers of her viewers etc.

We hypothesize that learning-by-viewing queries authored by analysts who outperform in both output-rate and social influence is associated with the largest improvement in productivity. To confirm the expert roles in learning-by-viewing, we start with testing the following hypothesis.

**Hypothesis 4.** *Past experience of viewing queries written by different types of data analysts (All-star, Maven, Lone-wolf or Non-star) is associated with different change in data analyst productivity.*

A rejection of Hypothesis 4 would imply that the predicted productivity improvement through viewing queries are independent of the type of the author. In contrast, support for Hypothesis 4 confirms the superiority of expert roles in our context and can be followed by further investigation: which type of analysts writes the most informative queries that are associated with largest productivity improvement of its viewers?

### 1.3. Methods

#### 1.3.1. Study Platform

We study eBay data analysts writing and viewing queries on *Alation*. *Alation* is an enterprise collaborative data platform developed by Alation Inc. and used by eBay Inc. As one of its clients. *Alation* serves as an all-inclusive ‘resort’ for data analysts. First, it provides a repository for all technical meta-data in the analytics data warehouse. Main
data services that can connect to *Alation* include Oracle, Teradata, MySQL, SQL Server and Tableau. A data analyst can conveniently access data if she has the proper permissions. Second, *Alation* integrates various analytics tools for data analysts to compose and execute queries, as well as produce comprehensible results. Third, *Alation* advances collaborations and social computing among data analysts inside eBay. A data analyst can share her knowledge with the community by publishing her queries, writing articles about her good practices or participating in conversations on technical issues. A data analyst can also seek knowledge from the community by viewing queries authored and published by other peers, searching for relevant articles or asking for help in the conversation board. Serving as an on-line enterprise community, *Alation* supports collaboration, knowledge sharing, reuse of resources, expertise location, innovation, organizational change and social networking (Matthews et al. 2014, 2013; Muller et al. 2012; Rowe et al. 2012); Everyone in the organization, from data novices to experts, can easily search, collaborate and leverage knowledge on *Alation*.

### 1.3.2. Empirical Setting

Our data consists of (1) the usage data of analysts on *Alation* which are automatically collected at the back end; and of (2) the employee information data that we scripted by crawling the eBay personal pages of all analysts in our study. The usage data include the following information spanning four years from January, 2014 to March, 2018: 1) records of entire queries on *Alation*, each item including query id, title, author, time of creation, a brief description of this query and whether this query has been published or not; 2) complete records of query executions on *Alation*, each item including query id,
user id, time of execution and number of statements that were executed; 3) complete records of all users viewing query pages on *Alation*; 4) simple personal information for all users on *Alation*, like username, email address, date of joining the *Alation* platform and date of last login. The employee information data track the public information of all data analysts in our study, including employee title, subsidiary area, manager path inside eBay, and location (city campus, country, and building and floor).

To construct our sample of data analysts, we first included all active users who have written at least one query on *Alation* during our study period. We then excluded users who are labeled as *Alation* employees and users who are authorized as *Alation* administrators inside eBay Inc. We also excluded users whose eBay employee information is missing on the eBay Intra-net. (That happens when an eBay employee left the company during the study period.) This resulted in an initial set of 2059 users. We then summarized users’ hierarchies inside eBay Inc. by parsing their manager paths. Among these 2059 users, there are 17 level-1 employees, 221 level-2, 777 level-3, 748 level-4, 281 level-5 and 15 level-6 (CEO is at level 10). We excluded level-6 and level-1 employees because they are either too senior or too inexperienced to be considered as representative data analysts in our study. The senior product manager who is in charge of *Alation* inside eBay Inc. also confirmed that level 2 - 5 employees are the major users. This resulted in a final data set of 2027 users that have written 101327 queries during the study period. We excluded queries that have never been executed during the study period nor exhibit missing field data; this finally left 79797 queries written by 2001 data analysts for the study. It is important to point out that only about 1 out of 8 queries is ever viewed by an analyst other than the author: out of the 79797 queries, only 10049
queries authored by 1097 data analysts have been viewed by analysts other than the authors.

1.3.3. Data and Measures

1.3.3.1. Dependent Variable. To develop a productivity measure for data analysts in our study, we borrow the concept of Pre-alpha phase from the software development life cycle (Tiwari 2010; Piggot and Amrit 2013; Buse and Weimer 2008). Figure 1.2 shows a complete development process as summarized by previous researchers (Rothfuss and Bauknecht 2002; Tiwari 2010). During the Pre-alpha phase developers write preliminary source code, which occurs before the Alpha testing.

Figure 1.3 illustrates the process that a data analyst follows when she is writing a query from scratch on Alation. We define FirstCompletionTime\text{analyst }i, \text{query }k as the time interval between the point when the data analyst \textit{i} clicks the button to create an empty query \textit{k} and the point when she executes this query for the first time:

\begin{equation}
\text{FirstCompletionTime}_{\text{analyst }i, \text{query }k} = \text{Timestamp}_{\text{first executes }k} - \text{Timestamp}_{\text{creates }k}
\end{equation}

This time interval, comparable to the Pre-alpha phase in software development, characterizes the time that a data analyst spends to shape her idea into a testable, prototype
Create an empty query on Alation

Start writing the code

Execute this query for the first time.

Not exactly what is wanted. Keep editing the code.

Execute

Not exactly what is wanted. Keep editing the code.

Execute

Not exactly what is wanted. Keep editing the code.

09:00:15 a.m. (Day 2)

It works perfectly!

Figure 1.3. The First Completion Time is the first stage of a typical query process.

query. The longer such time interval is, the later this analyst can proceed the following steps. We, therefore, use $FirstCompletionTime_{i,k}$ as a proxy for data analyst productivity. These query writing start time-stamps and the first execution time-stamps are automatically logged at the back end of Alation. We believe that the data analysts in our study cannot manipulate this data directly nor have no incentive to act strategically, as only the administrators of Alation are informed and have access to these data.  

1.3.3.2. Explanatory Variables.
Learning-by-doing. We define Aggregate Direct Experience$_{i,k}$ as the number of queries that a data analyst $i$ had written by herself on Alation, before she clicked the button to create an empty query $k$ during our study period. Because each query uses a particular database, we divide Aggregate Direct Experience$_{i,k}$ into two parts: 1) Direct Experience with the Focal Database$_{i,k}$, which is the number of queries using the same database as query $k$ uses, that $i$ has written by herself on Alation until she creates query $k$; 2) Direct Experience with Different Databases$_{i,k}$, which is the number of queries that $i$ has by herself written on Alation before she creates $k$ using a database different from the database $k$ uses. Clearly,

\begin{equation}
\text{Aggregate Direct Experience}_ {i,k} = \text{Direct Experience with the Focal Database}_ {i,k} + \text{Direct Experience with Different Databases}_ {i,k}
\end{equation}

Learning-by-viewing. We define Aggregate Indirect Experience$_{i,k}$ as the number of queries that a data analyst $i$ had viewed from her peers on Alation before she clicked the button to create an empty query $k$ during our study period. To differentiate learning-by-viewing from different types of data analysts, we further divide Aggregate Indirect Experience$_{i,k}$ based on types of the author analysts. For example, Indirect Experience from All-star$_{i,k}$ is the number of queries written by other all-star data analysts that $i$ had viewed before she clicked the button to create $k$. According to our segmentation of data analysts in figure 1.1 we divide Aggregate Indirect Experience$_{i,k}$ as in 1.5.
Aggregate Indirect Experience_{i,k} = \text{Indirect Experience from All-star}_{i,k} + \\
\text{Indirect Experience from Maven}_{i,k} + \\
\text{Indirect Experience from Lone-wolf}_{i,k} + \\
\text{Indirect Experience from Non-star}_{i,k}

(1.5)

To implement the analyst segmentation in Figure 1.1, we use Monthly Output of Queries on Alation as a measure of individual data analyst’s output-rate, and Monthly Viewers per Query as a measure of individual data analyst’s viewership, as defined earlier in equations (1.1–1.2). To calculate the PageRank, we build a directed, analyst-to-analyst network using the query-viewing data on Alation. Each node represents a data analyst in our study. The corresponding graph contains an edge from node A to node B if analyst A has viewed a query written by analyst B. Note that we weigh the edge from node A to node B using the number of distinct queries authored by B that analyst A has viewed. All self-loops have been excluded since we don’t consider the behavior that an analyst viewing her own queries as her social interaction. Running the PageRank algorithm implemented in the R package igraph returns the PageRank for all analysts (Csardi and Nepusz 2006; R Core Team 2018).

Figure 1.4 and 1.5 are scatter plots illustrating the relationship between output-rate and social influence. We apply the same segmentation model (see figure 1.1) to these two plots as follows: in the upper right quadrant we mark data analysts whose output-rate and social influence are both above the medians as all-star; in the lower left quadrant we mark data analysts whose output-rate and social influence are both below the medians.
Figure 1.4. Segmenting data analysts by Output-rate × Viewership

Note: The median is 0.67 for Output-rate and 0.06 for Viewership. The correlation between Output-rate and Viewership is 0.24. Points at the bottom of the plot with zero Viewership represent the data analysts whose queries haven’t been viewed by any peer; on a log scale these points fall at $-\infty$ but we moved them up for display convenience. These points heavily overlap because only about 1 out of 8 queries is ever viewed by a peer other than the author.

as non-star; we mark data analysts who reside in the upper left quadrant as maven and data analysts who reside in the lower right quadrant as lone-wolf.

1.3.3.3. Control Variables. Prior work suggests that individual adeptness, task complexity and multi-tasking are likely to affect productivity. A more adept data analyst may write a query faster; a data analyst who has piles of work may become less productive because of stress and pressure; a data analyst may spend more time on writing
Figure 1.5. Segmenting data analysts by Output-rate $\times$ PageRank

Note: The median is 0.67 for Output-rate and 0.0001 for PageRank. The correlation between Output-rate and PageRank is 0.16.

a complex query that either consists of many statements or uses a complicated database. We incorporate the following control variables to see if learning-by-doing and learning-by-viewing are associated with additive values over these factors in accelerating data-analyst query-writing. We explain the definitions of these control variables for the scenario in which the analyst $i$ creates and first executes query $k$.

- **Workload$_{i,k}$**: This is the average number of queries that analyst $i$ composes simultaneously together with the focal query $k$ during the $FirstCompletionTime_{i,k}$. This definition is inspired by the workload developed in Tan and Netessine 2014 for restaurant workers. For example, suppose $FirstCompletionTime_{i,k}$ lasts 40 minutes. During this period, the author data analyst only creates another
query that overlaps with the focal query $k$ for 20 minutes. The $Workload_{i,k}$, therefore, is $(40 \text{ min} + 20 \text{ min})/(40 \text{ min}) = 1.5$ queries.

- **Query Size$_k$:** This is the number of statements that were executed in the first execution of query $k$.

- **Database$_k$:** This is a categorical variable that indicates which database query $k$ uses.

- **Saved Query$_k$:** This is a binary variable that indicates whether query $k$ has been saved.

- **Migrated Query$_k$:** This is a binary variable that indicates whether part of query $k$ was migrated from a different platform. We acquire such information by parsing the title or description of query $k$.

- **Tenure on Alation$_{i,k}$: This is the number of months between the date when the author analyst $i$ joined Alation and the time she creates query $k$.

- **eBay Level$_i$:** This is a categorical variable that indicates author analyst $i$’s hierarchy in eBay Inc.

- **eBay Sub-area$_i$:** This is a categorical variable that indicates author analyst $i$’s subsidiary area in eBay Inc.

We also add month indicators for the number of months since January, 2014 until the creation of query $k$ to control for any environmental difference across time.

### 1.4. Analysis and Results

#### 1.4.1. Analysis Strategy
Table 1.1. Descriptive statistics of the raw data capturing $N = 79797$ queries written and executed by 2001 data analysts at eBay during 2014-2018.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Pctl(25)</th>
<th>Pctl(75)</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>FirstCompletionTime</td>
<td>90,711</td>
<td>1,200,935</td>
<td>1</td>
<td>20</td>
<td>262</td>
<td>83,798,568</td>
</tr>
<tr>
<td>Aggregate Direct Experience</td>
<td>116.2</td>
<td>200.2</td>
<td>0</td>
<td>16</td>
<td>126</td>
<td>1,775</td>
</tr>
<tr>
<td>Aggregate Indirect Experience</td>
<td>35.9</td>
<td>70.8</td>
<td>0</td>
<td>1</td>
<td>35</td>
<td>1,510</td>
</tr>
<tr>
<td>Direct Experience with Different Databases</td>
<td>52.4</td>
<td>124.8</td>
<td>0</td>
<td>2</td>
<td>49</td>
<td>1,389</td>
</tr>
<tr>
<td>Direct Experience with the focal Database</td>
<td>63.9</td>
<td>113</td>
<td>0</td>
<td>6</td>
<td>68</td>
<td>1,092</td>
</tr>
<tr>
<td>Workload</td>
<td>1.1</td>
<td>0.7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>Tenure on Alation</td>
<td>14.9</td>
<td>13.1</td>
<td>0</td>
<td>4</td>
<td>24</td>
<td>50</td>
</tr>
<tr>
<td>Saved Query</td>
<td>0.3</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Migrated Query</td>
<td>0.01</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Query Size</td>
<td>4.6</td>
<td>117.6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>18,095</td>
</tr>
<tr>
<td>Output-rate × Viewership Segmentation Indirect Experience from All-Star</td>
<td>32.6</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>1,343</td>
</tr>
</tbody>
</table>

Continued on next page
We present the descriptive statistics of all variables in the raw data in Table 1.1. Traditional learning curves are typically modeled as exponential forms and are often estimated using a log-linear regression model (Alchian 1963; Lapré et al. 2000; Argote 2012; Levy 1965). Our data are challenging in three ways: (1) our dependent variable

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Pctl(25)</th>
<th>Pctl(75)</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirect Experience from Lone-Wolf</td>
<td>0.2</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Indirect Experience from Maven</td>
<td>2.9</td>
<td>10.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>169</td>
</tr>
<tr>
<td>Indirect Experience from Non-star</td>
<td>0.1</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Indirect Experience from All-Star</td>
<td>32.8</td>
<td>64.4</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>1,352</td>
</tr>
<tr>
<td>Indirect Experience from Lone-wolf</td>
<td>0.01</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Indirect Experience from Maven</td>
<td>3</td>
<td>10.7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>169</td>
</tr>
<tr>
<td>Indirect Experience from Non-star</td>
<td>0.07</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
</tbody>
</table>
(FirstCompletionTime, in seconds) is a count variable and is over-dispersed, i.e., the variance is significantly larger than the mean (see table [1.1]); and (2) our dependent variable only takes positive values; (3) our data suffers from potential correlations across observations that are nested within multiple levels (a data analyst writes multiple queries using different databases over time). To deal with these challenges, we fit a mixed-effects zero-truncated negative binomial regression model. Zero-truncated negative binomial regression models are a class of generalized linear models that are appropriate for non-negative and over-dispersive count data (Allison and Waterman 2002). To account for the three-level nesting of the dataset (from queries to users to databases), we create a mixed-effects model where the experience variables and usage of databases are fixed effects and the unique analyst intercepts are represented as random effects (Karumur et al. 2018). In R (R Core Team 2018), mixed-effects zero-truncated negative binomial models are implemented in the glmmTMB package (Brooks et al. 2017).

For ease of comparing the relative importance of the explanatory variables, in the running models we standardize (i.e., normalize to mean zero and unit standard deviation) all variables except for the dependent variable and categorical variables.

1.4.2. Results

We build six separate models to test our hypotheses. Table [1.2] and [1.3] present the model specifications by indicating the inclusive explanatory variables in each model. Particularly, model 1 contains both Aggregate Direct Experience and Aggregate Indirect Experience; model 2 contains Direct Experience with the Focal Database, Direct Experience with Different Databases and Aggregate Indirect Experience. Model 3 and Model 4 are
Table 1.2. Results of Zero-truncated Negative Binomial Regressions for Model 1-2

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate direct experience</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td></td>
</tr>
<tr>
<td>Direct experience with focal database</td>
<td>-0.057*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td></td>
</tr>
<tr>
<td>Direct experience with different database</td>
<td>0.046*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td></td>
</tr>
<tr>
<td>Aggregate indirect experience</td>
<td>-0.029</td>
<td>-0.025</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.031)</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.36</td>
<td>-0.89</td>
</tr>
<tr>
<td></td>
<td>(24.39)</td>
<td>(19.93)</td>
</tr>
</tbody>
</table>

Note: * P<0.05; ** P<0.01; *** P<0.001. Control variables are omitted in the table.

built under the Output-rate × Viewership segmentation of data analysts: model 3 contains Aggregate Direct Experience and indirect experience respectively from four types of author analysts: all-star, non-star, maven and lone-wolf; model 4 contains Direct Experience with the Focal Database, Direct Experience with Different Databases, and indirect experience respectively from four types of author analysts. Except for being built under the Output-rate × PageRank segmentation, Model 5 and Model 6 have the same specifications as Model 3 and Model 4 have. All six models contain control variables that are listed in 3.3.3.
Table 1.3. Results of Zero-truncated Negative Binomial Regressions for Model 3-6

<table>
<thead>
<tr>
<th></th>
<th>Output-rate × Viewership Segmentation</th>
<th>Output-rate × PageRank Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Aggregate direct experience</td>
<td>−0.071</td>
<td>−0.045*</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.033)</td>
</tr>
<tr>
<td>Direct experience with focal database</td>
<td>−0.095***</td>
<td>−0.079***</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>Direct experience with diff. databases</td>
<td>−0.012</td>
<td>−0.003</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>Indirect experience: all-star</td>
<td>−0.144***</td>
<td>−0.143***</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Indirect experience: non-star</td>
<td>0.143***</td>
<td>0.151***</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.050)</td>
</tr>
<tr>
<td>Indirect experience: maven</td>
<td>0.277***</td>
<td>0.269***</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.034)</td>
</tr>
<tr>
<td>Indirect experience: lone-wolf</td>
<td>−0.182***</td>
<td>−0.178***</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.039)</td>
</tr>
<tr>
<td>Constant</td>
<td>−1.2</td>
<td>−0.74</td>
</tr>
<tr>
<td></td>
<td>(19.54)</td>
<td>(16.09)</td>
</tr>
</tbody>
</table>

Note: * P<0.05; ** P<0.01; *** P<0.001.
Control variables are omitted in the table.
Table 1.4. Comparison of Zero-truncated Negative Binomial Regression Models

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>−581,467</td>
<td>1,163,078</td>
<td>1,163,802</td>
</tr>
<tr>
<td>Model 2</td>
<td>−581,454</td>
<td>1,163,066</td>
<td>1,163,800</td>
</tr>
<tr>
<td>Model 3</td>
<td>−581,401</td>
<td>1,162,964</td>
<td>1,163,716</td>
</tr>
<tr>
<td>Model 4</td>
<td>−581,396</td>
<td>1,162,955</td>
<td>1,163,717</td>
</tr>
<tr>
<td>Model 5</td>
<td>−581,398</td>
<td>1,162,958</td>
<td>1,163,711</td>
</tr>
<tr>
<td>Model 6</td>
<td>(-581,393)</td>
<td>(1,162,950)</td>
<td>1,163,712</td>
</tr>
</tbody>
</table>

Note that model \{1, 2, 3, 4\} and model \{1, 2, 5, 6\} are nested: if we write Model 4 or 6 as

\[
(1.6) \quad g(\mathbb{E}[\text{FirstCompletionTime}_{i,k}]) = \beta_0 + \beta_1 \text{Direct Experience with Different DBs}_{i,k} + \\
\beta_2 \text{Direct Experience with the focal DB}_{i,k} + \\
\beta_3 \text{Indirect Experience from All-star}_{i,k} + \\
\beta_4 \text{Indirect Experience from Maven}_{i,k} + \\
\beta_5 \text{Indirect Experience from Lone-wolf}_{i,k} + \\
\beta_6 \text{Indirect Experience from Non-star}_{i,k} + \\
\gamma \text{Control Variables}_{i,k}
\]

where \(g(\cdot)\) is the general link function and is the natural logarithm in our model. Clearly, Model 1 is a special case of Model 4 or 6 (under different segmentations) where
\(\beta_1 = \beta_2\) and \(\beta_3 = \beta_4 = \beta_5 = \beta_6\); Model 2 is a special case of Model 4 or 6 where \(\beta_3 = \beta_4 = \beta_5 = \beta_6\); Model 3 is a special case of Model 4 where \(\beta_1 = \beta_2\). Model 5 is a special case of Model 6 where \(\beta_1 = \beta_2\). Similarly, Model 1 is a special case of Model 2 where \(\beta_1 = \beta_2\) and a special case of Model 3 where \(\beta_3 = \beta_4 = \beta_5 = \beta_6\); Model 2 is a special case of Model 4 or 6 where \(\beta_3 = \beta_4 = \beta_5 = \beta_6\) (under different segmentations).

We use Analysis of Variance (ANOVA) to test the nested models. Table 1.5 summarizes results of comparisons between nested pairs. We find that the difference in log-likelihoods of Model 1 and Model 2 is statistically significant. So is the difference in log-likelihoods of Model 1 and Model 3. Thus, we can reject both \(\beta_1 = \beta_2\) and \(\beta_3 = \beta_4 = \beta_5 = \beta_6\) (under output-rate \(\times\) viewership segmentation) with sufficient confidence. Other comparison results in Table 1.5 all support this finding. Similarly, the comparison results in Table 1.5 suggests that under output-rate \(\times\) PageRank characterization we can reject both \(\beta_1 = \beta_2\) and \(\beta_3 = \beta_4 = \beta_5 = \beta_6\).

1.4.2.1. Choosing the best model. The main results of the mixed-effects zero-truncated negative binomial regression on \FirstCompletionTime\ are reported in Table 1.2 and 1.3. Because we performed mean centering, our baseline was the mean values of all explanatory variables (except for the categorical ones). The coefficients can be interpreted as follows: for an explanatory variable change by one unit (in our case, by one standard deviation), the difference in the logs of expected counts of the dependent variable (\FirstCompletionTime) is expected to change by the corresponding coefficient, given all the other variables in the model are held constant.

We then use the Akaike Information Criterion (AIC) to evaluate the goodness of fit for each model. Generally, the smaller the AIC, the better the corresponding model.
Table 1.5. ANOVA results reject both $\beta_1 = \beta_2$ and $\beta_3 = \beta_4 = \beta_5 = \beta_6$

<table>
<thead>
<tr>
<th></th>
<th>$\Delta$Df</th>
<th>$\chi^2$</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1: Model 2</td>
<td>1</td>
<td>14</td>
<td>&lt; 0.001***</td>
</tr>
<tr>
<td>Model 1: Model 3</td>
<td>3</td>
<td>119.64</td>
<td>&lt; 0.001***</td>
</tr>
<tr>
<td>Model 1: Model 5</td>
<td>3</td>
<td>125.31</td>
<td>&lt; 0.001***</td>
</tr>
<tr>
<td>Model 1: Model 4</td>
<td>4</td>
<td>130.62</td>
<td>&lt; 0.001***</td>
</tr>
<tr>
<td>Model 1: Model 6</td>
<td>4</td>
<td>135.48</td>
<td>&lt; 0.001***</td>
</tr>
<tr>
<td>Model 2: Model 4</td>
<td>3</td>
<td>116.26</td>
<td>&lt; 0.001***</td>
</tr>
<tr>
<td>Model 2: Model 6</td>
<td>3</td>
<td>121.48</td>
<td>&lt; 0.001***</td>
</tr>
<tr>
<td>Model 3: Model 4</td>
<td>1</td>
<td>10.625</td>
<td>0.001**</td>
</tr>
<tr>
<td>Model 5: Model 6</td>
<td>1</td>
<td>10.171</td>
<td>0.001**</td>
</tr>
</tbody>
</table>

over other competing models. Under this criterion, Model 4 is the best among the four models $\{1, 2, 3, 4\}$ that apply the output-rate $\times$ viewership segmentation of data analysts; Model 6 is the best among the four models $\{1, 2, 5, 6\}$ that apply the output-rate $\times$ PageRank segmentation of data analysts. We thereby adopt Model 4 and Model 6 to understand learning-by-doing and learning-by-viewing on data analysts in our study, under the respective segmentation of data analysts. To avoid multicollinearity between explanatory variables in these two final models, we examine the VIF (variance inflation factor) of the set of explanatory variables, comparing against the recommended maximum of 5 (Cohen et al. 2014). Table 1.6 shows in our case the VIFs of all explanatory variables in the final models remain well below 2, indicating the absence of multicollinearity.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Output-rate × Viewership Segmentation</th>
<th>Output-rate × PageRank Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct experience with the focal database</td>
<td>1.28</td>
<td>1.29</td>
</tr>
<tr>
<td>Direct experience with different databases</td>
<td>1.29</td>
<td>1.29</td>
</tr>
<tr>
<td>Indirect experience from All-Star</td>
<td>1.52</td>
<td>1.50</td>
</tr>
<tr>
<td>Indirect experience from Lone-Wolf</td>
<td>1.30</td>
<td>1.17</td>
</tr>
<tr>
<td>Indirect experience from Maven</td>
<td>1.58</td>
<td>1.62</td>
</tr>
<tr>
<td>Indirect experience from Non-Star</td>
<td>1.32</td>
<td>1.36</td>
</tr>
<tr>
<td>Workload</td>
<td>1.02</td>
<td>1.04</td>
</tr>
<tr>
<td>Tenure on Alation</td>
<td>1.04</td>
<td>1.05</td>
</tr>
<tr>
<td>Saved Query</td>
<td>1.05</td>
<td>1.05</td>
</tr>
<tr>
<td>Migrated Query</td>
<td>1.04</td>
<td>1.05</td>
</tr>
<tr>
<td>Query Size</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>Mean VIF</td>
<td>1.26</td>
<td>1.23</td>
</tr>
</tbody>
</table>

**1.4.2.2. Output-rate × viewership segmentation.** From Model 4 in Table 1.3 we find that analysts who have more prior direct experience with the focal database are more likely to have shorter expected *FirstCompletionTime* of writing new queries. Holding all other variables constant, a unit (in our case, by one standard deviation) increase in a data analyst’s direct experience with the focal database is significantly associated
with an average 9.5% decrease in the expected *FirstCompletionTime* of a new query. Estimated by the mean of *FirstCompletionTime*, 9.5% decrease is equivalent to 2.4 hours less. A unit increase in direct experience with different databases predicts a 1.2% shorter *FirstCompletionTime*, yet not statistically significant.

We also find that viewing queries authored by different types analysts predicts different changes in the expected *FirstCompletionTime* of new queries. Holding all other variables constant, a unit increase in a data analyst’s indirect experience (the number of queries she has viewed) from all-star analysts is significantly associated with an average 14.3% decrease (equivalent to 3.6 hours less if estimated by the mean) in the expected time she would spend between creating and first executing a new query; a unit increase in the number of queries the focal data analyst has viewed from lone-wolf is significantly associated with an average 17.8% decrease (equivalent to 4.4 hours less if estimated by the mean). Both indirect experience from non-star and maven is associated with a significant increase in the expected *FirstCompletionTime* of a new query.

1.4.2.3. **Output-rate × PageRank characterization.** From Model 6 in Table 1.3 we find that more prior direct experience with the focal database is associated with a shorter expected *FirstCompletionTime* of new queries. Holding all other variables constant, a unit increase in a data analyst’s direct experience with the focal database is significantly associated with an average 7.9% decrease (equivalent to 2.0 hours less if estimated by the mean) in the expected *FirstCompletionTime* of a new query; a unit increase in her direct experience with different databases is linked with the decrease that is neither statistically nor piratically significant (0.3% on average).
We also find that viewing queries authored by different types of analysts is associated with different changes in the expected $\textit{FirstCompletionTime}$ of a new query. Holding all other variables constant, a unit increase in a data analyst’s indirect experience from all-star analysts is significantly associated with an average 19.6% decrease (equivalent to 4.9 hours less if estimated by the mean) in the expected $\textit{FirstCompletionTime}$ of a new query; a unit increase in the number of queries the focal data analyst has viewed from lone-wolf authors is significantly associated with an average 17.1% decrease (equivalent to 4.3 hours less if estimated by the mean). Both indirect experience from non-star and maven are related with a significant increase in the expected $\textit{FirstCompletionTime}$ of a new query.

To summarize, our results provide robust support for hypothesis 2 and 4 under both segmentations of data analysts. We have partial support for hypothesis 1 because our results suggest only the direct experience with the focal database is associated with significant improvement in data-analyst productivity. We also have partial support for hypothesis 3 because our results suggest that only viewing queries authored by all-star and lone-wolf analysts is associated with significant improvement in data analyst productivity. Analysts who have viewed more queries authored by maven and non-star analysts are more likely to spend longer $\textit{FirstCompletionTime}$ on a new query in our study.

Furthermore, we find that under the $\textit{Output-rate} \times \textit{PageRank}$ segmentation, viewing queries authored by all-star analysts is associated with the largest improvement
in data-analyst productivity. In contrast, under the Output-rate × viewership segmentation, viewing queries authored by lone-wolf analysts is associated with the largest improvement.

1.5. Summary, Discussion, and Limitations

1.5.1. Summary and Discussion

Our results provide evidence of a statistically-significant association between both learning-by-doing and learning-by-viewing and data-analyst productivity. Greater direct experience with the focal database and greater indirect experience from viewing queries authored by ‘all-star’ and ‘lone-wolf’ analysts both predict significant improvement in data analyst’s productivity. The magnitude of the coefficients in table 1.3 underscore the practical significance of these associated value. For instance, an increase by one standard deviation in a data analyst’s direct experience with the focal database predicts a significant decrease of 9.5% in the expected FirstCompletionTime, the equivalent of 2.4 hours average reduction. One standard deviation increase in the number of queries a data analyst has viewed from ‘all-star’ authors is associated with a significant decrease of 14.3% in the expected FirstCompletionTime, under the output-rate × viewership characterization, or a significant decrease of 19.6% under the output-rate × PageRank characterization. These are equivalent to a reduction of 3.6 or 4.9 hours respectively in the expected FirstCompletionTime. With 1600 queries created on Alation every month in our study period, these numbers accumulate to substantial numbers.

Our study builds on and contributes to the literature of organizational learning and computational social science, as well as that of cognitive psychology literature
and learning sciences. Although the relationship between experience and productivity have been extensively studied in manufacturing and service industries, there are only few empirical studies of learning processes among data analysts and none with this granularity of field data. Most existing studies are qualitative studies, such as interviews and surveys, and tend to measure learning using cognitive approaches (Kandel et al. 2012; Kim et al. 2017; Brehmer et al. 2014; Kim et al. 2016; Kandogan et al. 2014). To the best of our knowledge, our study may be the first to empirically examine how data analysts learn to speed up writing queries using behavioral (performance-measure) approaches.

First, our results support association between learning from direct experience and productivity. More experience in writing queries is associated with faster FirstCompletionTime. This finding is consistent with the study by Kim et al. that surveyed 793 professional data scientists at Microsoft (Kim et al. 2017). Respondents to the survey spoke of ‘getting their hands dirty’ as one of the best practices to improve data science analysis, and frequently mentioned the desire for hands-on training and practical case studies. We also find that prior experience in using the focal database is associated with greater improvement in productivity than the experience of using a different database. This implies that data analysts obtain more domain knowledge of a specific database through self-practice (Kang and Hahn 2009). This finding also echoes previous studies that demonstrate the greater impact of related experience on productivity (Kc and Staats 2012; Staats and Gino 2012; Egelman et al. 2016; Wong 2004).
Second, our results provide evidence for the role of “experts” in learning-by-viewing. We find that only viewing queries authored by analysts with high output rate is associated with significant improvement in data analyst productivity. This finding not only confirms that circulating the query code can be an effective social practice, but also suggests that interacting with the ‘community of practices’ is critical for the advance of an analyst’s knowledge. This is consistent with the finding of Dabbish et al. that being able to view code authored by others supports better programming (Dabbish et al. 2012). Kim et al. also reported that “respondents expressed the goal of fostering a ‘community of practice’ across the company” (Kim et al. 2017). Despite of this, Kandel et al. found in their interview study that “the least commonly shared resource among data analysts was the analysis code” (Kandel et al. 2012). Our results give support to the value of collaborative platforms in providing an unfenced channel for collaboration.

Third, our findings suggest that different types of social influence, namely the ‘local’ influence (as captured by the viewership) vs. the ‘global’ influence (as captured by the page-rank), are associated with different changes in viewer analysts’ productivity. Under both segmentations of analysts, only viewing queries authored by ‘all-star’ and ‘lone-wolf’ is associated with a significant decrease in the FirstCompletionTime of new queries. Analysts who mostly queries authored by ‘non-star’ and ‘maven’ analysts, in contrast, are more likely to spend more of that time. Using the output-rate × viewership segmentation, we find that viewing queries authored by ‘lone-wolf’ analysts is associated with the largest improvement in FirstCompletionTime; using the output-rate
× page-rank segmentation, we find that viewing queries authored by ‘all-star’ analysts is associated with the largest improvement. These two findings together indicate that the most influential analysts might be those that have few incoming links (fewer direct viewers) but a relatively large contingency of viewers with high PageRank. Such analysts could be the ‘ultimate stars’.

Overall, regardless of the segmentation of analysts, learning-by-viewing is associated with greater productivity improvement than learning-by-doing in our study. This result agrees with Lave and co-authors’ conclusion that ‘engaging in practice may well be a condition for the effectiveness of learning’ and is consistent with the anecdotal evidence in their paper that apprentices learn mostly in relations with other apprentices (Lave et al. 1991).

1.5.2. Implications

While one must be careful with extrapolations, our results suggest directions for further explorations by managers and designers.

1.5.2.1. Managerial Implications. First, our findings about learning-by-doing provide guidance for developing training plans for data analysts. Based on the result that learning-by-doing with the same data bases is associated with significant productivity improvement, managers can encourage data analysts to practice with focus. Similarly, individual contributors on peer production community (e.g. Wikipedia and GitHub) can choose to work on related projects in order to be more productive.
Second, our findings about learning-by-viewing provide suggestive evidence of the value of collaborative platforms. Furthermore, the differential associations of learning-by-viewing queries authored by different types of analysts can inform performance evaluation and team-building. Organizations that use collaborative platforms can take their employees’ star status on such platforms seriously into evaluation of their performance and contribution. Besides, with the identification of star vs. non-star, project managers can team suitable analysts to work together.

Third, our findings help to identify star workers on collaborative platforms. We show it is useful to leverage articulated social measures and observed code-related activity simultaneously. As Marlow et al. put it (Marlow et al. 2013), “Succinctly summarizing expertise based on behavioral data and incorporating evidence of social interactions may support more nuanced impressions and reduce bias. In any type of peer production site where a person shares their work for others to build on, dealing with contributions from others is necessary and important.” Collaborative platforms or peer production communities may consider adopting measurements such as viewership and PageRank that are discussed in our study or other measurements that fit their context better.

1.5.2.2. Design Implications. Our results particularly inform the future design of collaborative platforms to support learning. We emphasize providing cues about expertise and star status, by showing that a data analyst learns better if she has more interactions with certain groups of star analysts. Accessible cues about expertise and star status can be very critical for the legitimate periphery participation of the beginners or novice (Lave et al. 1991). Researches in corporate domain also suggest that expertise
finding is an important task (e.g. (Raban et al. 2012)) and show that many internal tools have been developed to help people tag their own and others’ expertise (Shami et al. 2009). Our results suggest that collaborative platforms invest in designing tools to deliberately highlight certain information thereby providing social signals. For example, dashboard pages and activity feeds could signal to data analysts the social significance and technical merits of their peers. The theory of social translucence also confirms the potential for this transparency to radically improve collaboration and learning in complex knowledge-based activities (Dabbish et al. 2012). For learners, analysts or developers can take these cues into effective strategies for coordinating projects and advancing their technical skills (Erickson and Kellogg 2000). For knowledge contributors, signaling their social influence among the communities may motivate desired behavior yet one always should anticipate unintended consequences.

1.5.3. Limitations and Future Work

First, our study focuses on the behavioral approach to measure learning. Qualitative studies that describe the cognitive phenomena of analysts are lacking. Central questions like, “what are the analysts learning” or “are they really learning anything” remain to be answered directly. Future work to interview, observe, or survey data analysts will bring insightful answers to these questions. These follow-up qualitative studies asking data analysts to explain their behaviors in creating/viewing queries could complement our findings towards a better understanding of the mechanisms underlying these two learning processes. For example, researchers who find the negative association between viewing queries authored by ‘maven’ and the viewer’s productivity
rather intriguing can survey the subgroup of analysts in our study who have viewed most queries from ‘mavens’. Perhaps these analysts found it very difficult to match their existing knowledge with those queries authored by ‘maven’ (Von Mayrhauser and Vans 1995). Furthermore, we only report the analyst’s activity of writing/viewing a query without further reporting which part of the query that the focal analyst has indeed focused on. Such additional information may help us identify what domain-specific aspects of viewing peers’ queries or what type of self-practice advance learning efficiency more (Fritz et al. 2007, 2014). Still this requires more qualitative evidence in the future work.

Second, our results show associations without supporting causal inferences. Although we apply mixed-effects models to address potential correlations and incorporate control variables to deal with some confounding factors, our analysis may still suffer from endogeneity issues such as self-selection bias. Nonetheless, the associations we find do motivate randomized experiments in future research to carefully measure causal effects. One possible way to do this is to bring exogenous variations in analysts’ exposure to queries authored by peers. Randomizing query-searching results and arbitrarily mask peers’ queries to analysts are both practical.

Third, we use FirstCompletionTime to measure data-analyst productivity, which is the period of time from a data analyst creating an empty query to executing it for the first time. Other (aggregate) metrics, like the number of queries a data analyst creates in a week or month, may highly depend on the assignment that the analysts has received. Nevertheless, it is possible that FirstCompletionTime depends on individual work style. Some empirical evidence in the learning sciences shows two distinct programming
styles: tinkering vs. planning (Turkle and Papert 1992). Tinkerers keep exploring new ideas by making adjustments step by step. Planners first develop a clear plan, then do it once and right (Resnick and Rosenbaum 2013). Tinkering data analysts may be making small and frequent code changes and thus have shorter FirstCompletionTime than planning analysts. Such caveat may be addressed by using a different measure of the coding time: the period of time that the data analyst remains active on the composing page. To obtain such time requires detailed click-stream data that is not accessible to us.

Fourth, our study focuses exclusively on the quantity (productivity) without discussing the quality of queries. Including the quality of a query is not straightforward as different people may have different opinions on what constitutes a good query: teammates or project managers may want the query to be well-documented so that it can be reused; IT staff may want the query to be efficient when running on large-scale data; stakeholders or decision-makers may want the query to deliver accurate and insightful answers to their questions; the author may want the query to run smoothly without errors. Again, to operationalize these multiple perspectives on quality requires different levels of data, both qualitative and quantitative, which are not accessible in our data. Future research could extend our study by generating measures of quality (e.g., the number of errors during executions, or the rate of positive responses from peers who have viewed the query) to examine the results with quality-adjusted output.
1.6. Conclusion

We investigate in this paper two learning processes—learning-by-doing and learning-by-viewing—on data analysts. We position our models within two modes of organizational learning (learning from direct experience vs. learning from indirect experience) and build our hypotheses on literature of organization science and learning sciences. We pose and empirically test several hypotheses in the context of eBay data analysts writing and viewing queries on *Alation*. We find that learning-by-doing is associated with significant improvement in data analyst productivity when the corresponding prior experience is with the focal database. We also find that the association between learning-by-viewing and productivity improvement depends on the type of the author analyst: Only viewing queries authored by ‘all-star’ and ‘lone-wolf’ data analysts is associated with significant productivity improvement in data analyst writing new query. Our results also suggest that different measures of author analysts’ social influence, namely the ‘local’ influence captured by viewership vs. the ‘global’ influence captured by *PageRank*, relate with different changes in viewer analysts’ productivity. Using the output-rate $\times$ viewership characterization, viewing queries authored by ‘lone-wolf’ analysts (i.e. analysts with higher output-rate and less viewership) was associated with the highest improvement; using the output-rate $\times$ page-rank characterization, viewing queries authored by ‘all-star’ analysts (i.e. analysts with both higher output-rate and higher PageRank) was associated with the highest improvement. These two findings suggest that the ‘ultimate stars’ may be those with with higher output-rate and lower ‘local’ influence (i.e. less viewership) yet higher ‘global’ influence (i.e. higher PageRank).
CHAPTER 2

A Hidden Markov Model of Data Analyst Productivity and Learning Dynamics

(Joint with Jan A. Van Mieghem and Itai Gurvich)

2.1. Introduction

A well-developed literature in economics and management studies learning curve within a broad spectrum of industries from manufacturing to services. Benkard (1999) builds a learning model for a commercial aircraft manufacturer in which the unit production cost depends on a firm’s past production experience. Darr et al. (1995) demonstrate that, in pizza franchises, the unit cost to produce a pizza decreases significantly with direct experience in production. More recent studies of learning curve focus on knowledge-intensive industries such as IT consulting and software development (Fong Boh et al. 2007; Kim et al. 2012; Kang and Hahn 2009). Most of these studies assume a static model of the learning effect in which performance changes at a uniform and constant rate with cumulative experience. However, assuming this type of static learning behavior for all units (individual, group or organization) could be unsatisfying. First, not all units learn at the same rate. Second, even for the same unit, its learning rate may vary over time. It has been demonstrated that experience on one learning task may influence and improve performance on some subsequent learning (Ellis 1965). In other words, when a unit is equipped with sufficient knowledge stock
and expertise, accumulating the same kind of experience may have a different effect on improving its performance than previously. This paper presents a model to capture such learning dynamics at individual level. Specifically, we propose a Hidden Markov Model (HMM) to study the heterogeneous evolvement of productivity and learning ability among a group of data analysts, using their historical participation in several learning activities.

We focus on two modes of learning activities: writing queries and viewing peer’s queries. Analysts’ improvement from writing queries is a typical example of learning-by-doing, i.e. analysts can learn from their direct experience (of writing queries). In contrast, their improvement from viewing queries written by their peers is an instance of learning from indirect experience. Learning from direct experience and from indirect experience are two distinctive modes of learning that have been well studied in the stream of learning-curve literature (Reagans et al. 2005; Fong Boh et al. 2007; Kc et al. 2013). Extensive empirical evidence of learning from direct experience and indirect experience has been found in various organizations (Darr et al. 1995; Staats and Gino 2012; Kim et al. 2012; Argote 2012; Darr et al. 1995; Staats 2012; Huckman et al. 2009; Reagans et al. 2016; Huckman and Staats 2011).

Participation in these learning activities builds up the knowledge stock and expertise of analysts, thereby affecting their long-term performance and subsequent assimilation of new information. We characterize an individual data analyst in each time period by a state that corresponds to specific productivity and learning ability. The proposed HMM consists of a finite set of such states and allows a dynamic membership of analysts to these states. The transitions between the states are determined by
the participation in the two learning activities. We also incorporate a set of analyst-specific random effects in the HMM to control over unobserved heterogeneity in learning and writing queries at individual level. The HMM is estimated using a Markov chain Monte Carlo (MCMC) hierarchical Bayes procedure.

The rest of the paper is organized as follows. In section 2.2, we present the literature and theoretical background. In section 2.3 we develop the HMM to explain the learning of data analysts. Section 2.4 illustrates the empirical setting, including the data and variables. The estimation procedure and results are presented in section 2.5 and 2.6. Section 2.7 offers discussion on future work and concluding remarks.

2.2. Literature and Theory Development

2.2.1. Individual Learning from Direct Experience

The concept of learning curve for individuals has been widely studied in psychology for over one century. The first individual learning curve is demonstrated by Ebbinghaus (1885) to characterize individual memorizing ever-longer strings on nonsense syllables. Wright (1936) provided evidence that the production cost of airframe declined with increasing cumulative output. “Improvement in proficiency of a workman with practice”, the first explanation proposed by Wright for the evidence he observed, formally characterizes individual learning curve in manufacturing. Such process that individuals create knowledge from the experience directly acquired by themselves is referred to as learning from direct experience (Argote 2012). Recent studies of individual learning from direct experience cover various industries. Fong Boh et al. (2007) provide empirical findings in software development by revealing that specialized direct experience
has the greatest impact on the productivity of individual developers. Staats and Gino (2012) demonstrate the positive impact of task repetition on individual worker productivity in a Japanese bank. Kim et al. (2012) show that IT consultants demonstrate a distinct learning curve on average resolution time as they accumulate individual problem-solving experience. Kc and Staats (2012) examine the direct impact of individual cardiologist’s own prior experience on their performance. Clark et al. (2013) also find evidence of learning from direct experience for individual providers of outsourced services, by showing the benefits of customer-specific experience accumulated by individual radiologists. Individual developers in open source software projects are also found to benefit from their own experience (Singh et al. 2011).

2.2.2. Individual Learning from Indirect Experience

Prior research on organizational learning confirms that individuals also benefit from their peers’ experience, which is referred to as learning from indirect experience (Gino et al. 2010; Huber 1991; Levitt and March 1988; Miner and Haunschild 1995). The situative perspective of learning also suggests that the individual mental process cannot completely account for learning. It is now well accepted that learning is frequently an interactive, social phenomenon (Tyre and Von Hippel 1997). Such communal processes are important in organization because no one person has sufficient knowledge necessary for solving all complex problems. It has been shown in the context of machine repair technicians, that knowledge does not come from what is taught in the classroom, but rather from informal story-sharing among technicians and users about their experiences in particular work environments (Brown and Duguid 2000; Orr 2016).
Similar evidence in the context of computer programming is provided by [Brandt et al. (2009)]. By relying on information and source code fragments provided by other people from the Web, developers engage in just-in-time learning of new skills and approaches, clarify and extend their existing knowledge, and remind themselves of details deemed not worth remembering. [Vasilescu et al. (2013)] argue that participating in on-line programming community (e.g. StackOverflow) speeds up development activities as quick solutions to technical challenges are provided by peers, thus saving the developers precious time. Notwithstanding, current empirical evidence for the benefit of learning from indirect experience is inconclusive. [Waldinger (2011)] finds no evidence for peer effects on the productivity of researchers in physics, chemistry and mathematics. In his study, even very high-quality scientists do not affect the productivity of their local peers.

2.2.3. Motivation of Dynamic Learning

Most existing learning research, including that we have mentioned above, adopts a static model: the learning occurs with cumulative experience at a constant rate, and individual’s ability (i.e. unobservable factors that contribute to individual’s performance) is stable. However, it is possible that past experience will be accumulated as "human capital", thereby changing individual learning ability and influencing future learning. [Cohen and Levinthal (1990)] refer to the ability to evaluate and utilize outside knowledge as absorptive ability, and argue that such ability is largely a function of the level of prior related knowledge. Studies in the area of cognitive and behavior science at individual level also justify and enrich the observation that prior knowledge
or experience on learning task may influence subsequent learning (Ellis 1965; Lindsay and Norman 2013; Bower et al. 1981). Under this circumstance, sticking to the static model might lead to erroneous estimates of the impact of participating in learning activities. Manski (1993) also proposes that static model of peer effects can suffer from the reflection problem by assuming that individual performance is influenced by the endogenously aggregated group performance. Reflection problem challenges the identification of peer effects and may cause spuriously over-estimated learning effects. One potential solution he suggests to this problem is to apply dynamic models. But only a few empirical studies have incorporated dynamics in learning models. Chan et al. (2014) investigate the long-term effect of peer-based learning on individual salesperson’s productivity. In their model, these effects build up, increase the salesperson’s ability and thereby change the efficiency of learning over time. Singh et al. (2011) develops a stochastic model to capture developer learning dynamics in open source software projects.

2.2.4. Hidden Markov Model for Data Analysts Learning

The latent nature and structure of such dynamics make them very difficult to capture. To analyze the learning dynamics of data analysts in our setting, we propose a Hidden Markov Model (HMM) similar to that in Singh et al. (2011). The HMM is a model of bivariate Markov process \( \{S_t, O_t\}_{t=1}^{\infty} \) where \( \{S_t\}_{t=1}^{\infty} \) is a Markov process that consists of a finite state space and \( \{O_t\}_{t=1}^{\infty} \) is a sequence of observations that is determined by a state-specific stochastic process. Generally \( \{S_t\}_{t=1}^{\infty} \) is assumed not directly observable (hidden). The possibility of modeling complex structures of \( \{O_t\}_{t=1}^{\infty} \) through a
simple formulation makes HMMs attractive and widely applied in areas ranging from bioinformatics (Leroux and Puterman 1992), ecology (Morf 1998) and criminology, to speech recognition (Rabiner and Juang 1986) and transportation (MacDonald and Zucchini 1997; Scott 2002). They are also found to be very important in econometrics and macro-economics. For example, Hamilton (1989) proposed an HMM as a very tractable approach to modeling unobservable regime shifts and to estimating the impact of such shifts on the level of U.S. real gross national product. Another well-known tool in the quantitative finance literature with essentially the same structure of HMM is the Kalman filter (Koopman 1997; Duan and Simonato 1999).

The HMM model described in this section is an individual-level model of query writing and learning behaviors. In our HMM, we characterize these hidden states with respect to the latent productivity and learning ability of data analysts, or absorptive ability in Cohen and Levinthal (1990)’s term. Such latent characteristics can include programming skills, knowledge on the databases and other factors that may influence analyst productivity and subsequent learning. At any point of time, a data analyst resides in only one state. The transitions between states over a period are affected by her participation in learning activities. Such activities can include the data analyst’s writing her own queries (i.e. learning by doing) and her viewing on queries written by other analysts (i.e. learning by viewing). The relationship of a state with a unique productivity behavior in writing query, along with the Markovian transitions between states allow us to capture data analyst’s productivity and learning dynamics.

To do so, we use a nonhomogeneous HMM (Hughes et al. 1999) in which the Markovian transitions are a function of time-varying variables. This is important for us to
understand the drivers of the dynamics rather than merely building a model that fits
the dynamics in the data. Specifically, we first assume that a data analyst’s probability
for moving from one state to another is determined by her observed participation in
learning activities as well as the unobserved individual random effects. Second, the
observation in our HMM is the productivity of individual analyst in writing a query
in different periods of time. The query-specific productivity is a function of analyst’s
intrinsic productivity, observed characteristics of the query, and unobserved individ-
ual random effects. Adding random effects to both parts of the model grants us the
possibility to capture the heterogeneity of analysts’ performance in both learning and
writing queries.

Figure 2.1 shows how our HMM relates the transitions between latent states to ob-
served dynamics in productivity. Specifically, in period \( t \), an analyst probabilistically
resides in one of total \( N_s \) states based on her accumulated learning stock. Each realized
state corresponds to a unique stochastic process that characterizes the analyst’s pro-
ductivity in writing queries in period \( t \). Her participation in learning activities during
period \( t \) determines whether she will move up to a higher state, stay where she is, or
move down to a lower state in the next period.

This model has several advantages compared with the standard learning curve
model. First, it allows us to capture the learning dynamics of a data analyst by in-
vestigating the differential impact of learning activities on her transitions between dif-
ferent hidden states, conditional on the state she currently resides. Second, this model
also provides us a dynamic segmentation of data analysts based on their time-varying
states, whereas an ad-hoc specification of is usually added to an otherwise static model.
Third, the use of state and the inherent dependency between serial states within the model enable us to fully account the effect of autocorrelation. Autocorrelation in our setting may occur when an unobservable shock (e.g. unexpected illness of the individual analyst) commonly affects serial or multiple query-writing observations. This unobservable shock, if only exists in one period, can be captured by the state-specific query-writing process that is common to all observations in that period. If the shock spans across two serial periods, the state that corresponds to the previous period can
carry on the effect of the unobservable shock to the current state thanks to the Markov-
ian dependency between serial states. Fourth, in this model the current query-writing
behavior only depends on the lag participation in learning activities. By doing so it
diminishes reverse causality between learning and change in productivity. An analyst
may view many queries from peers because she needs to write an extremely complex
query. Such reverse causality may lead to erroneous estimate of learning effects in the
classical static model.

2.3. Model Development

Consider a finite set of states \( s \in \{1, 2, ..., N_s\} \) in our HMM, with 1 being the lowest
state and \( N_s \) being the highest. The entire time horizon is divided in to a finite num-
ber of periods \( t \in \{1, 2, ..., T\} \). At the beginning of any given period, a data analyst
resides in an unknown state. For each period and each data analyst, we observe the
productivity of the data analyst in writing queries, other factors that can affect her pro-
ductivity and her participation in various learning activities during the period. The
observed sequence of a data analyst \( i \)'s productivity over periods is denoted as her
outcome sequence \( O^i = \{O^i_1, O^i_2, ..., O^i_T\} \), where \( O^i_t \) characterizes her temporal produc-
tivity in writing queries in period \( t \). Given these primitives, our HMM has three main
components:

i The initial state distribution, \( \pi \);

ii The sequence of the state-transition probability matrix, \( Q \), with \( Q_{t-1 \rightarrow t} \) denoting
the Markovian transition matrix from \( t - 1 \) to \( t \);
iii The observed productivity with their state-dependent probability $P$. The probability that data analyst $i$ is observed with certain productivity at time $t$ conditional on her state is $P(O_i^t|S_i^t)$, where $S_i^t$ is the state that $i$ belongs to in period $t$ and $O_i^t$ is the observed productivity of $i$ in the same period.

2.3.1. The Initial State Distribution

The initial state distribution characterizes the probability that an average data analyst starts in a particular state at the beginning of time horizon (period $t = 1$). Let $\pi_s$ be the probability that an average analyst belongs to the state $s$ in period $t = 1$, where $s \in \{1, 2, ..., n\}$. We have $\sum_{s=1,2,...,N_s} \pi_s = 1$. In our HMM, we assume initially the probabilistic membership of all data analysts to different states follows the same distribution:

\begin{equation}
\pi = (\pi_1, \pi_2, ..., \pi_{N_s})
\end{equation}

Many studies assuming time-homogeneous HMM define the initial state distribution as the stationary distribution of the transition matrix (MacDonald and Zucchini 1997; Netzer et al. 2008; Montoya and Raffaelli 2010). However, in our HMM the transition matrix is time-varying, thereby violating the assumption of homogeneity. Some other studies estimate their HMMs with fixed initial state distribution. Such assumption can come from prior information (Schweidel et al. 2011) or is simply specified for the computational convenience (Singh et al. 2011). An alternative shown in a recent study by Lee et al. (2017) is to directly estimate the initial state probability distribution at the
population level. In our HMM we fix the initial state distribution to be uniform across states, i.e. \( \pi_s = \frac{1}{N_s}, \forall s \in \{1, 2, ..., N_s\} \).

### 2.3.2. The State Transition Matrix

We model the transitions between latent learning states as a Markov process where only transitions to adjacent states are feasible. The transition matrix is defined as the follows:

\[
Q_{t \rightarrow t+1}^i = \begin{bmatrix}
q_{1,1}^i & q_{1,2}^i & 0 & \cdots & 0 & 0 \\
q_{2,1}^i & q_{2,2}^i & q_{2,3}^i & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & q_{N_s, N_s-1}^i & q_{N_s, N_s}^i
\end{bmatrix}
\]

where \( q_{s,s'}^i = P(S_{t+1}^i = s'|S_t^i = s) \) is the conditional probability that analyst \( i \) moves from state \( s \) in period \( t \) to state \( s' \) in period \( t \), and \( 0 \leq q_{s,s'}^i \leq 1 \ \forall s,s', \ \sum_{s'=1,2,\ldots,N_s} q_{s,s'}^i = 1 \ \forall s \). We model the transitions between the states as a threshold model; what transition occurs depends on how the stock of learning compares to the threshold. The stock of learning, termed as \( LS_s^i \), is assumed to be a function of analyst \( i \)'s participation in learning activities during the period as the follows:

\[
LS_s^{it} = \beta_s \text{Learning activities}_{it} + \zeta_i + \epsilon_{ist}
\]
where Learning activities$_{it}$ is vector of $i$’s participation in learning activities in this period and $\beta_s$ is a vector of coefficients capturing the effect of participation in these activities on building up the analyst’s learning stock. Note that the effects of participation in learning activities are state-specific. Researchers in program cognition models have shown that past experiences and knowledge base of a programmer greatly affect her program understanding, thereby affecting the return of learning (Von Mayrhauser and Vans 1995). $\zeta_i$ is the random effect that captures the unobserved heterogeneity in state transitions across analysts. Such unobserved heterogeneity may include participation in offline learning activities that we do not observe. Assuming the unobserved $e_{ist}$ is independently and identically distributed (IID) of the extreme value type, the non-homogeneous transition probabilities are modeled as an ordered logit (Greene and Hensher 2010):

\[
\begin{align*}
q_{it}^{s,s+1} &= 1 - \frac{\exp(\mu(s + 1, s) - \beta_s \text{Learning activities}_{it} - \zeta_i)}{1 + \exp(\mu(s + 1, s) - \beta_s \text{Learning activities}_{it} - \zeta_i)} \\
q_{it}^{s,s-1} &= \frac{\exp(\mu(s - 1, s) - \beta_s \text{Learning activities}_{it} - \zeta_i)}{1 + \exp(\mu(s - 1, s) - \beta_s \text{Learning activities}_{it} - \zeta_i)} \\
q_{it}^{s,s} &= \frac{\exp(\mu(s + 1, s) - \beta_s \text{Learning activities}_{it} - \zeta_i)}{1 + \exp(\mu(s + 1, s) - \beta_s \text{Learning activities}_{it} - \zeta_i)} - \frac{\exp(\mu(s - 1, s) - \beta_s \text{Learning activities}_{it} - \zeta_i)}{1 + \exp(\mu(s - 1, s) - \beta_s \text{Learning activities}_{it} - \zeta_i)}
\end{align*}
\] (2.4)

where $\mu(s + 1, s)$ is the logit threshold for analysts in state $s$ to move up to state $s + 1$ and $\mu(s - 1, s)$ the logit threshold for analysts in state $s$ to move down to state $s - 1$. Put it into words, a data analyst will move up to a higher state if her stock of learning during the period is greater than $\mu(s + 1, s)$, or move down to a lower state if her
temporal learning is smaller than $\mu(s - 1, s)$. We constrain $\mu(s + 1, s) > \mu(s - 1, s), \forall s \in \{1, 2, ..., n\}$ to ensure the order of these thresholds. We also fix $\mu(N_s + 1, n) = \infty$ and $\mu(0, 1) = -\infty$ to ensure that analysts already in the lowest or highest state transit correctly. Analysts of states $s \in \{2, 3, ..., N_s - 1\}$ can move up to a higher state, move down to a lower state or stay in the current state.

2.3.3. State-Dependent Productivity in Writing Queries

We use $FirstCompletionTime_{it}^{kt}$ as a proxy for data analyst productivity. For the detailed definition of $FirstCompletionTime_{it}^{kt}$, see 1.3.3.1. Specifically, it is the time interval between the point when the data analyst $i$ clicks the button to create an empty query $k$ and the point when she executes this query for the first time, in period $t$:

$$\text{(2.5) } FirstCompletionTime_{it}^{kt} = \text{Timestamp}_{i \text{ first executes } k} - \text{Timestamp}_{i \text{ creates the empty } k}$$

Because $FirstCompletionTime_{it}^{kt}$ is a count variable and suffers from over-dispersion, i.e. the variance is significantly larger than the mean, we assume it follows a state-dependent negative binomial distribution:

$$\text{(2.6) } P(FirstCompletionTime_{it}^{kt} = \tau | S_i = s) = \frac{e^{-\lambda_{itk}} \lambda_{itk}^\tau}{\Gamma(\tau + 1)}$$

where the conditional mean $\lambda_{itk}$ is defined as the following:

$$\text{(2.7) } \log(\lambda_{itk}) = \rho_s Z_{itk} + \eta_i + \epsilon_{itk}$$
where $Z_{itk}$ is a vector of time-varying factors that can affect the productivity of analyst $i$ in writing query $k$ in period $t$. $\rho_s$ is a vector of state-specific coefficients that capture the effect of $Z_{itk}$, where $s$ is the state that $i$ resides in period $t$. The random effect $\eta_i$ captures the analyst-specific unobserved heterogeneity, which may include individual programming style (Blikstein et al. 2014). With the log-gamma distribution of the error term $\epsilon_{itk}$, we can reformulate equation 2.6 as the following (Greene 2011):

\[
(2.8) \quad P(\text{FirstCompletionTime}_{itk}^i = \tau | S_i^t = s) = \frac{\Gamma(\delta_s + \tau)}{\Gamma(\tau + 1)\Gamma(\delta_s)} h_{itk}^{\delta_s}(1 - h_{itk})^\tau
\]

where

\[
(2.9) \quad h_{itk} = \frac{\delta_s}{e^{\rho_s Z_{itk} + \eta_i} + \delta_s}
\]

with $\delta_s > 0$ as a state-specific dispersion parameter.

Note that in our empirical setting, a data analyst can write multiple queries in one period. That is, given a data analyst $i$ has written $K$ queries in period $t$, the productivity outcome we can observe in the period is a sequence of the FirstCompletionTime as the follows:

\[
(2.10) \quad O_i^t = \{\text{FirstCompletionTime}_{1i}^t, \text{FirstCompletionTime}_{2i}^t, \ldots, \text{FirstCompletionTime}_{Ki}^t\}
\]

Therefore, the probability of observing $O_i^t$ is:

\[
(2.11) \quad P(O_i^t | S_i^t = s) = \prod_{k=1,2,\ldots,K} P(\text{FirstCompletionTime}_{itk}^i | S_i^t = s)
\]
2.3.4. The Likelihood of an Observed Sequence of Productivity Outcome

Because of the Markovian structure of this model, the probabilities of an individual’s productivity outcome are correlated through the underlying sequence of the hidden states. Specifically, consider the realized state path for an analyst \( i \) is \( S^i = \{ S^i_1, S^i_2, ..., S^i_T \} \) and the observed sequence of outcome \( O^i = \{ O^i_1, O^i_2, ..., O^i_T \} \). Then we have

\[
P(O^i | S^i) = \prod_{t=1}^{T} P(O^i_t | S^i_t)
\]

And the probability of seeing such state path is:

\[
P(S^i) = P(S^i_1) \prod_{t=1}^{T-1} q_{S^i_t S^i_{t+1}}^t
\]

where \( P(S^i_1) \) is the probability that the initial state of \( i \) is \( S^i_1 \); here \( P(S^i_1 = s) = \pi_s \). So the joint probability of \( S^i \) and \( O^i \) is

\[
P(S^i, O^i) = P(S^i_1) \prod_{t=1}^{T-1} q_{S^i_t S^i_{t+1}}^t \prod_{t=1}^{T} P(O^i_t | S^i_t)
\]

Accordingly, the likelihood of observing the sequence of \( O^i \) is given by summing equation 2.14 over all possible state paths that \( i \) could take overtime:

\[
L(O^i) = \sum_{S^i_1=1}^{n} \sum_{S^i_2=1}^{n} ... \sum_{S^i_T=1}^{n} P(S^i, O^i)
\]

Following MacDonald and Zucchini (1997), we can simplify equation 2.15 into the following matrix notation:

\[
L(O^i) = \pi P^i_1 Q^i_{1 \rightarrow 2} P^i_2 Q^i_{2 \rightarrow 3} ... P^i_T 1'
\]
where

\[
P_t^i = \begin{pmatrix}
P(O_i^i | S_i^i = 1) & 0 & 0 & \cdots & 0 & 0 \\
0 & P(O_i^i | S_i^i = 2) & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & P(O_i^i | S_i^i = N_s)
\end{pmatrix}
\]

and \(1'\) is an \(N_s \times 1\) vector of ones. The full likelihood across analysts is simply the production of individual likelihood over \(i \in \{1, 2, ..., N_{\text{ind}}\}\).

2.4. Empirical Setting

2.4.1. Study Platform and Data

We study eBay data analysts writing and viewing queries on Alation. See 1.3.1 and 1.3.2. for more details about the study platform and our construction of data.

2.4.2. Variables

Because we have explained our dependent variable, \(\text{FirstCompletionTime}_{ik}\) in section 2.3.3., we only describe variables related with analyst’s participation in learning activities (i.e. Learning activities\(_{it}\) and variables related with analyst’s productivity in writing query (i.e. \(Z_{itk}\) in Equation 2.7).

2.4.2.1. Variables Related with Learning Activities.

- \(\text{Number of Queries Written}_{it}\): this is defined as the number of queries that a data analyst \(i\) has written by herself on Alation during period \(t\). The mean of \(\text{Number of Queries Written}_{it}\) is 0.782 and the standard deviation is 3.82.
• Number of Queries Viewed \(_{it}\): this is defined as the number of queries that a data analyst \(i\) has viewed from her peers on Alation during period \(t\). The mean of Number of Queries Viewed \(_{it}\) is 0.306 and the standard deviation is 2.03.

2.4.2.2. Variables Related with Productivity in Writing Queries. We continue the usage of Workload\(_{i,k}\), Query Size\(_k\), Saved Query\(_k\), Migrated Query\(_k\), Tenure on Alation\(_{i,k}\) as variables that influence the productivity of analyst \(i\) in writing query \(k\). See 1.3.3.3. for the definition of these variables. The descriptive statistics for these variables can be found in Table [1.1]. In addition to these variables, we include a state-specific constant in \(Z_{itk}\). This constant term captures a state-specific fixed effect in query-writing productivity, which can be interpreted as the intrinsic productivity in writing queries for an average analyst in the corresponding state.

2.5. Estimation Strategy

Two main approaches have been proposed to estimate the parameters of an HMM: (1) maximum likelihood estimation either by direct optimization or by Expectation Maximization algorithm [Dempster et al. 1977; Baum 1972] and (2) Bayes estimation. We estimate our HMM using a standard hierarchical Bayes estimation procedure described in [Rossi et al. 2012], for its great flexibility and efficiency in estimating models that account for cross-sectional heterogeneity. [Heckman 1981] illustrated the importance of properly accounting for heterogeneity in dynamic models by showing that ignoring heterogeneity might lead to an spurious overstatement of the estimated effects.
Instead of focusing on the point estimate of true parameter value as the frequentist approach, the main interest of the Bayes approach is in generating the entire distribution of the parameters given the data and a prior. Bayes theorem provides the mechanism for how data and prior beliefs are translated into posterior beliefs:

\[
p(\theta | y) = \frac{p(y | \theta) p(\theta)}{p(y)} \propto p(y | \theta) p(\theta)
\]

where \( \theta \) is the parameters we want to estimate and \( p(\theta | y) \) is the posterior distribution of \( \theta \), given observed data \( y \) and prior \( p(\theta) \). Inference here refers to summarizing the posterior distribution with typical statistics, such as the posterior mean \( E[\theta] = \int \theta p(\theta | y) d\theta \) and posterior standard deviation. Note that Bayesian inference here is conducted only using formal probability theory and does not require asymptotic approximations, which is particularly useful in modeling data with limited size.

To apply a standard hierarchical Bayes estimation procedure, we distinguish two sets of parameters: random-effect parameters \( \{\Theta_i\} \) and parameters that are common across individuals \( \Psi \). Specifically,

\[
\Theta_i = \{\zeta_i, \eta_i\}
\]

\[
\Psi = \{\mu(2, 1), \mu(1, 2), \mu(3, 2), ..., \mu(s - 1, s), \mu(s + 1, s), ..., \mu(N_s - 1, N_s),
\]

\[
\delta_1, \delta_2, ..., \delta_{N_s}, \beta_1, \beta_2, ..., \beta_s, ..., \beta_{N_s},
\]

\[
\rho_1, \rho_2, ..., \rho_s, ..., \rho_{N_s}\}
\]
and we assume a prior distribution for the random parameters as the following:

\[
\text{Prior}(\Theta_i) = \text{Normal}(0, \Sigma_{\Theta})
\]

Conforming to convention, we complete this specification by assuming the typical diffuse priors for \(\Psi\) and the hyperparameter \(\Sigma_{\theta}\), following Netzer et al. (2008):

\[
\text{Prior}(\Sigma_{\Theta}) = \text{Inverse Wishart}(N_{\Theta} + 5 + N_{\text{ind}}, I_{N_{\Theta}})
\]

\[
\text{Prior}(\Psi) = \text{Normal}(0, 30I_{N_{\Psi}})
\]

Conducting bayes estimation requires sequential drawing from the conditional posterior \(p(\{\Theta_i\}, \Psi|y)\) using the Markov Chain Monte Carlo (MCMC) algorithm. Note that Equation 2.18 along with Equation 2.17 suggests that the conditional posterior does not have a closed form. Thus, we apply Metropolis-Hasting algorithm (Metropolis et al. 1953; Hastings 1970) to sample from the posterior distribution, using a Gibbs sampler Geman and Geman (1984).

The basis of the Metropolis algorithm consists of: (1) in the \(m^{th}\) iteration, simulate a candidate sample of parameters, say \(\Omega^{cand}\), from the proposal distribution \(\Phi^{m-1}(\cdot)\) based on the last accepted sample \(\Omega^{m-1}\); (2) accept the candidate sample with the probability calculated via the acceptance function \(\alpha(\Omega^{cand} | \Omega^{m-1})\). We choose the Gaussian distributions as the proposal distributions, taking advantage of its symmetry. That is \(\Phi^{m-1}(\cdot) = \text{Normal}(\Omega^{m-1}, \Sigma)\). The typical acceptance function is the following:

\[
\alpha(\Omega^{cand} | \Omega^{m-1}) = \min\{1, \frac{P(\Omega^{cand}|y)}{P(\Omega^{m-1}|y)}\}
\]
where $P(\cdot | y)$ is the posterior joint density given the data. This acceptance function is designed to send the sampler to move to higher probability areas under the posterior joint density. In practice, we draw a random number uniformly between 0 and 1, and compare it with $\frac{P(\Omega^{\text{cand}} | y)}{P(\Omega^{m-1} | y)}$. If $\frac{P(\Omega^{\text{cand}} | y)}{P(\Omega^{m-1} | y)}$ is larger than the uniform random number, we accept the candidate sample; otherwise we reject it. After running the algorithm for a long time, we end up with a sequential sample from the target posterior distribution of the parameters.

Gibbs sampler is an MCMC sampler developed by Geman and Geman (1984), which allows us to divide variables of interest into blocks consisting of two or more variables and sample the joint distribution of each block conditional on all other variables. Gibbs sampler is particularly suitable for hierarchical model. With Gibbs sampler, we can recursively update the individual random effects $\{\Theta_i\}$, the covariance $\Sigma_\theta$ and the common effects $\Psi$ following Algorithm I.

### 2.6. Results

We estimate two models of a single state and of three states respectively. The one state model assumes contemporaneous effect of writing and viewing queries on analyst productivity. It is equivalent to the classic learning curve model. We use the maximum likelihood procedure to estimate this model. The three-state model is estimated within the MCMC hierarchical Bayes procedure, using the Gibbs sampler and the random-walk Metropolis-Hastings algorithm implemented in Julia. We ran the MCMC with random initial values of 100,000 iterations until the convergence
Algorithm 1: Hierarchical Bayes Estimation Algorithm

Initialize: Generate a random sample of
\[ \Sigma_0 \sim \text{Inverse Wishart}(N_\Theta + 5 + N_{ind}, I_{N_\Theta}) \]
\[ \{\Theta_0^i\} \sim \text{Normal}(0, \Sigma_0) \]
\[ \Psi_0 \sim \text{Normal}(0, 30I_{N_\Psi}) \]

for iteration \( m = 1, 2, \ldots \) do

Step 1

for \( i \leftarrow 1 \) to \( N_{ind} \) do

1. Simulate a candidate \( \Theta_i^{\text{cand}} \sim \text{Normal}(\Theta_i^{m-1}, \lambda_\Theta \bar{\Sigma}_\Theta) \), where \( \Theta_i^{m-1} \) is the accepted draw in iteration \( m - 1 \), and \( \lambda_\Theta \) and \( \bar{\Sigma}_\Theta \) are chosen adaptively following the Algorithm 4 in [Andrieu and Thoms (2008) and Shaby and Wells (2010)]. We adapt \( \lambda_\Theta \) and \( \bar{\Sigma}_\Theta \) to reduce the autocorrelation between draws and to maintain the optimal acceptance rate.

2. Accept \( \Theta_i^{\text{cand}} \) to be \( \Theta_i^m \) with the probability of

\[
\min \left\{ \frac{\exp(-\frac{1}{2} \Theta_i^{\text{cand}} (\Sigma_\Theta^{m-1})^{-1} \Theta_i^{\text{cand}}) P_i(y_i|\Theta_i^{\text{cand}}, \Psi_i^{m-1})}{\exp(-\frac{1}{2} \Theta_i^{m-1} (\Sigma_\Theta^{m-1})^{-1} \Theta_i^{m-1}) P_i(y_i|\Theta_i^{m-1}, \Psi_i^{m-1})}, 1 \right\}
\]

if reject, \( \Theta_i^m = \Theta_i^{m-1} \).

Step 2

Sample \( \Sigma_\Theta^m \sim \text{Inverse Wishart}(N_\Theta + 5 + N_{ind}, I_{N_\Theta} + \sum_{i=1}^{N_{ind}} \Theta_i^m \Theta_i^{m'}) \)

Step 3

Given the updated \( \{\Theta_i^m\} \)

1. Simulate a candidate \( \Psi^{\text{cand}} \sim \text{Normal}(\Psi^{m-1}, \lambda_\Psi \bar{\Sigma}_\Psi) \), where \( \Psi^{m-1} \) is the accepted draw in iteration \( m - 1 \). \( \lambda_\Psi \) and \( \bar{\Sigma}_\Psi \) are also adaptively chosen using the same algorithm as in Step 1.

2. Accept \( \Psi^{\text{cand}} \) as \( \Psi^m \) with the probability of

\[
\min \left\{ \frac{\exp(-\frac{1}{2} \Psi^{\text{cand}} (V_0^{m-1})^{-1} \Psi^{\text{cand}}) P_i(y_i|\Theta_i^m, \Psi^{\text{cand}})}{\exp(-\frac{1}{2} \Psi^{m-1} (V_0^{m-1})^{-1} \Psi^{m-1}) P_i(y_i|\Theta_i^m, \Psi^{m-1})}, 1 \right\}
\]

where \( V_0^m = 30I_{N_\Psi} \). If reject, \( \Psi^m = \Psi^{m-1} \).

is achieved, with two chains. Convergence was assessed by comparing the within-variance to the between-variance for each parameter estimated across chains ([Brooks and Gelman 1998]). The first 85,000 iterations serve as "burn-in" iterations and only the last 15,000 was kept for the inference. Table 2.1 shows that the one-state model is
Table 2.1. Fit Measures for Comparing Models

<table>
<thead>
<tr>
<th>Model</th>
<th>-2LogLikelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-state Model</td>
<td>1 232 380.8</td>
<td>1 232 400.8</td>
<td>1 232 493.6</td>
</tr>
<tr>
<td>Three-state Model</td>
<td>433 477.1</td>
<td>441 551.1</td>
<td>479 043.7</td>
</tr>
</tbody>
</table>

Note: The more negative the number, the better the fit. AIC, Akaike information criterion; BIC, Bayesian information criterion.

outperformed by the three-state model in all the relevant fit measures. This implies the advantage of accounting for the learning dynamics of analysts.

2.6.1. HMM Estimates

Table 2.2 presents the summary statistics, including means, standard deviation and 95% highest posterior density (HPD) intervals, calculated from the posterior distribution for all parameters. The estimates that have non-zero-overlapping HPDs are referred to as "significant".

2.6.1.1. State Dependent Productivity. We notice that there is a clear variation in the state-specific intrinsic productivity of analysts, which is indicated by the constant term $\rho_{\text{constant}}^1$, $\rho_{\text{constant}}^2$ and $\rho_{\text{constant}}^3$. The posterior means are 5.8, 3.2 and $-6.8$ for state 1, 2, 3 respectively, and all of them are significant. This clearly indicates that the intrinsic productivity of analysts improves with the rank of states they belong to (more negative constant implies less FirstCompletingTime). This pattern grants us one potential interpretation of these states: analysts in state 1 seem to be in an inexperienced state and write queries in such a slack manner, we label state 1 as novice. Accordingly, in ascending order, we label state 2 as intermediate and state 3 as advanced, since analysts in latter states seem more capable and have higher productivity in writing queries.
Table 2.2. Estimation Results for the Three-state Hidden Markov Models Parameter

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior mean</th>
<th>Posterior Standard Deviation</th>
<th>95% HPD intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{1}^{\text{writing}}$</td>
<td>5.9</td>
<td>1.4</td>
<td>(4.5, 8.4)</td>
</tr>
<tr>
<td>$\beta_{2}^{\text{writing}}$</td>
<td>4.4</td>
<td>1.8</td>
<td>(2.5, 7.5)</td>
</tr>
<tr>
<td>$\beta_{3}^{\text{writing}}$</td>
<td>0.3</td>
<td>0.4</td>
<td>(−0.1, 1.0)</td>
</tr>
<tr>
<td>$\beta_{1}^{\text{viewing}}$</td>
<td>5.3</td>
<td>1.2</td>
<td>(4.0, 7.0)</td>
</tr>
<tr>
<td>$\beta_{2}^{\text{viewing}}$</td>
<td>−8.4</td>
<td>0.8</td>
<td>(−10.0, −7.5)</td>
</tr>
<tr>
<td>$\beta_{3}^{\text{viewing}}$</td>
<td>−8.3</td>
<td>1.4</td>
<td>(−10.4, −6.5)</td>
</tr>
<tr>
<td>$\mu(2, 1)$</td>
<td>20.5</td>
<td>1.6</td>
<td>(17.2, 22.6)</td>
</tr>
<tr>
<td>$\mu(2, 3)$</td>
<td>−3.5</td>
<td>2.9</td>
<td>(−7.0, 0.4)</td>
</tr>
<tr>
<td>$\mu(1, 2)$</td>
<td>−13.3</td>
<td>2.0</td>
<td>(−16.6, −11.2)</td>
</tr>
<tr>
<td>$\log(\mu(3, 2) - \mu(1, 2))$</td>
<td></td>
<td>13.0</td>
<td>(11.4, 14.0)</td>
</tr>
<tr>
<td>$\log\delta_{1}$</td>
<td>−1.8 × 10^{-4}</td>
<td>1.8 × 10^{-4}</td>
<td>(−6.7 × 10^{-4}, 4.7 × 10^{-6})</td>
</tr>
<tr>
<td>$\log\delta_{2}$</td>
<td>10.8</td>
<td>0.8</td>
<td>(9.8, 12.0)</td>
</tr>
<tr>
<td>$\log\delta_{3}$</td>
<td>−2.3</td>
<td>0.2</td>
<td>(−2.8, −2.1)</td>
</tr>
<tr>
<td>$\rho_{\text{constant}}^{2}$</td>
<td>5.8</td>
<td>2.4</td>
<td>(3.7, 10.0)</td>
</tr>
<tr>
<td>$\rho_{\text{constant}}^{3}$</td>
<td>3.2</td>
<td>0.6</td>
<td>(2.5, 4.0)</td>
</tr>
<tr>
<td>$\rho_{\text{constant}}^{\text{workload}}$</td>
<td>9.9</td>
<td>0.5</td>
<td>(8.9, 10.8)</td>
</tr>
<tr>
<td>$\rho_{\text{workload}}^{2}$</td>
<td>2.8</td>
<td>2.3</td>
<td>(−0.1, 6.6)</td>
</tr>
<tr>
<td>$\rho_{\text{workload}}^{3}$</td>
<td>7.7</td>
<td>1.2</td>
<td>(5.3, 8.7)</td>
</tr>
<tr>
<td>$\rho_{\text{tenure of author}}^{1}$</td>
<td>6.3 × 10^{-3}</td>
<td>0.1</td>
<td>(−0.1, 0.1)</td>
</tr>
<tr>
<td>$\rho_{\text{tenure of author}}^{2}$</td>
<td>−8.1</td>
<td>1.8</td>
<td>(−11.6, −6.1)</td>
</tr>
<tr>
<td>$\rho_{\text{tenure of author}}^{3}$</td>
<td>−9.0</td>
<td>1.8</td>
<td>(−11.1, −6.6)</td>
</tr>
<tr>
<td>$\rho_{\text{migrated query}}^{1}$</td>
<td>2.3</td>
<td>3.2</td>
<td>(−1.4, 6.8)</td>
</tr>
<tr>
<td>$\rho_{\text{migrated query}}^{2}$</td>
<td>−4.6</td>
<td>1.8</td>
<td>(−7.0, −2.3)</td>
</tr>
<tr>
<td>$\rho_{\text{migrated query}}^{3}$</td>
<td>9.2</td>
<td>0.4</td>
<td>(8.3, 9.7)</td>
</tr>
<tr>
<td>$\rho_{\text{saved query}}^{1}$</td>
<td>0.3</td>
<td>0.1</td>
<td>(0.2, 0.5)</td>
</tr>
<tr>
<td>$\rho_{\text{saved query}}^{2}$</td>
<td>8.4</td>
<td>2.9</td>
<td>(4.8, 12.7)</td>
</tr>
<tr>
<td>$\rho_{\text{saved query}}^{3}$</td>
<td>−1.1 × 10^{-2}</td>
<td>0.6</td>
<td>(−1.0, 1.2)</td>
</tr>
<tr>
<td>$\rho_{\text{query size}}^{1}$</td>
<td>−3.5</td>
<td>2.4</td>
<td>(−8.6, 0.0)</td>
</tr>
<tr>
<td>$\rho_{\text{query size}}^{2}$</td>
<td>−4.8</td>
<td>2.5</td>
<td>(−9.4, 0.6)</td>
</tr>
<tr>
<td>$\rho_{\text{query size}}^{3}$</td>
<td>0.2</td>
<td>0.3</td>
<td>(−0.1, 1.0)</td>
</tr>
</tbody>
</table>

Note: Bolded coefficients’ HPDs do not cross zero.
The posterior means for the effects of workload are 9.9, 2.8 and 7.7 respectively. The positive estimates in each state indicate that higher workload an analyst takes on will decrease her productivity in writing queries. Notice that this effect is insignificant and the smallest in the magnitude on analysts in the intermediate state.

The posterior means for the effects of analysts’ tenure on Alation are 0.006, −8.1 and 9.0 respectively. This indicates that analysts who are in intermediate state and above benefit from their time involved with Alation. For analysts in the novice state this effect is rather insignificant.

Our results also show the relationship between some query-specific characteristics with the FirstCompletionTime. A migrated query was originally created in other platform and later moved by analysts to Alation. Our results show that analysts in novice state are insensitive to whether the current query is migrated or not. On contrary, there is significant decrease in the time that analysts in intermediate state spent on migrated queries (prior mean is -4.6 and is significant), whereas a significant slowdown is observed on analysts in advanced state when they are working on the migrated queries. One potential explanation for this pattern is analysts in advanced state are usually asked to migrate large and complex queries. Ensuring those queries work appropriately on a new platform requires more time and caution as more calibration and comments may be necessary. Our results also suggest significant increase in the time that analysts in novice and intermediate states spent on queries that had been be saved. Saving a query may signal the complexity of the query, which requires longer time to finish. But this effect is insignificant for analysts in advanced state. We observe no significant effects of the query size on the FirstCompletionTime in any state. This is
probably because most (more than 75%) of the queries in our study only consist of one statement.

2.6.1.2. Learning and State Transitions. The threshold for moving from novice up to intermediate state is 20.5, and that for moving from advanced state down to intermediate is $-3.5$. For analysts in intermediate state, the threshold is $-13.3$ to move down to the novice state and $422932.8$ to move up to the advanced state. This indicates that it gets more difficult to move to a higher state from the intermediate state than from the novice state. The parameters corresponding to learning activities are all significant, except for learning from writing own queries in advanced state. Our results show both writing own queries and viewing peers’ queries seem to help the novice analysts to move up to the intermediate state. The effect of learning from writing own queries remains positive though smaller for analysts in intermediate state. This finding is supported by Shafer et al. (2001), who have found that people with higher expertise in a given area have lower subsequent learning rate in that area. We also observe that analysts in intermediate state and above risk moving down to a lower state by viewing peers’ queries. One potential reason is that the quality of published queries may be very uneven as suggested in Chapter 1. Here we only consider the aggregate views for an analyst without differentiating the source of knowledge as we do in Chapter 1. Also, looking at peers’ codes—albeit appealing—may delay programming activities, since both of these activities compete for the time resources of developers. Analysts in higher states may find most materials, albeit consistently time-consuming, as helpful as to novice analysts. Indeed, it is well known that “a wealth of information creates
a poverty of attention and a need to allocate that attention efficiently among the over-
abundance of information sources that might consume it” (Anderson et al. 1997). An-
other potential explanation is the switch of domain: when an analyst in intermediate
state or above are observed to view queries more often, it is highly possible that she
starts working on something new, such as a brand new business unit or an unfamiliar
database. In this case we can expect her later performance to be more like a novice.

2.7. Concluding Remarks

Classical literature of individual learning curve assumes a static model, which may
lead to biased estimates of learning effects if the individual behavior evolves dynam-
ically. In this paper we presents a Hidden Markov Model to capture such dynamics
in the productivity and learning of data analysts. We compare our model with the
standard learning curve framework. The results indicate that the HMM is more ap-
propriate. The proposed HMM allows us to label three ordered states that analysts can
move up through learning and are found with increasing intrinsic productivity of an-
alysts. We also examine the effects of participating in two modes of learning activities,
writing own queries (learning from direct experience) versus viewing peers’ queries (learning
from indirect experience), on the analysts’ transition between these states. Our results
suggest differential effects across the current learning states of analysts.

Several managerial implications can be drawn from our study. First, using our
model managers can identify the learning states of analysts at any time. Knowing
the learning state of an analyst means knowing her intrinsic productivity and learn-
ing ability. Such information can be used for job assignment to optimize temporary
productivity. Once a state is identified, the manager can motivate appropriate learning activity based on our results to advocate the transition to and sustention in higher states. Our findings reveal that novice analysts benefit from both writing and viewing queries while intermediate analysts only benefit from writing queries. Taking this into consideration can help managers design more efficient training programs for analysts to move up to higher states. To sustain analysts in the advanced state, the managers need to divert them from viewing peers’ queries to producing more queries on their own. Furthermore, our model can be used by managers to predict when an initially peripheral analyst may evolve into an advanced analyst through participation in appropriate learning activities. Knowing this adds rich information on the cost-benefit analysis for recruiting. If a newbie can soon grow into a proficient analyst with appropriate learning, managers may not need to offer high salary to attract a veteran.

We highlight several limitations and directions that we plan to explore in the next stage. First, in this study we only estimate one-state and three-state HMMs. Although the findings confirm that the dynamic model is more appropriate for our empirical setting, the three-state HMM may not be the best-fitting model. We should estimate multiple HMMs with different number of states to find the optimal number of states. Second, the uniform assumption of initial states could be too strong. A more direct modeling approach we plan to explore is to estimate the initial state probability distribution at the population level (Lee et al. [2017]). Finally, we can extend the hierarchical model by allowing individual random effects to be a function of observed individual characteristics, such as demographics, locations, organizational hierarchy, etc (Allenby and Ginter [1995]; Netzer et al. [2008]).
CHAPTER 3

Resident Supervision and Patient Care: A Comparative Time Study in a Community-Academic Versus a Community Emergency Department

(Joint with Ernest E. Wang, Jan A. Van Mieghem and Itai Gurvich)

3.1. Introduction

3.1.1. Background

Emergency Departments (EDs) can broadly be categorized into three types: academic - defined as university-based, teaching hospitals, community-academic (which will be referenced as “CAED” in the remainder of this paper) defined as community EDs with residents and/or students who rotate through and are supervised by Emergency Physicians (EPs), and non-academic EDs (which will be referred as “community ED” in the remainder of this paper) defined as an ED with no learners. At CAED, resident supervision is but one of many responsibilities that must be skillfully orchestrated by EPs, alongside other essential tasks such as direct patient care, communication, and

This chapter is a reprint: Ernest E Wang, Yue Yin, Itai Gurvich, Morris S Kharasch, Clifford Rice, Jared Novack, Christine Babcock, James Ahn, Steven H Bowman, and Jan A Van Mieghem. Resident supervision and patient care: A comparative time study in a community-academic versus a community emergency department. AEM Education and Training, 2019. License Number: 4645720632438
documentation. One of the primary benefits of residency training is the opportunity for residents to observe the work of, and learn from the direct interaction with, the attending emergency physician (EP) \cite{Merritt2017}. Direct attending-resident interaction is also considered as the best assessment tools for resident competencies \cite{Chisholm2004}.

### 3.1.2. Importance

The optimal teacher-learner interaction in residency education entails listening to the learners, asking questions, and leading the learners to conclusions rather than supplying them. Maintaining effective supervision can be very costly for emergency departments (ED). First, engaging with residents may limit the time that EPs devote to patient care and ED flow. This added demand on the EP may also increase stress and anxiety, potentially leading to burnout. Second, the EP is responsible for calibrating the level of supervision based on each resident’s knowledge and clinical skill. Residents who lack requisite knowledge with complex cases may struggle and potentially cause patient harm \cite{Atzema2005}. Third, supervising residents consume resources. Evidence shows that care at academic hospitals is less cost-effective than care at non-academic hospitals because of higher frequency of testing and other resource use in teaching setting \cite{Pitts2014}.

In 2001, DeBehnke wrote, "Educating and supervising residents and students while simultaneously providing patient care requires quantifiable faculty time and effort. Academic EDs must identify this time and effort accurately since providing this joint
product line has the potential to make our emergency care system inefficient. Notwithstanding, limited data has been reported previously that identifies the time and effort spent by EPs at CAEDs in comparison to community EDs. Chisholm et al. tracked EPs’s time expenditures on direct/indirect patient care and personal activities in academic versus community EDs, but did not specify the time spent supervising residents or performing other care-related activities (Chisholm et al. 2004, 2011). Other studies assessed aggregate resident effect on departmental throughput (Svirsky et al. 2013; Bhat et al. 2014; McGarry et al. 2010; Anderson et al. 2013; Henning et al. 2013). To the best of our knowledge, this is the first study to comprehensively quantify and compare the time EPs spend on resident supervision and care-related activities in CAEDs versus community EDs.

3.1.3. Objective

The objective of this study was to compare the time utilization profiles of a group of EPs in a community ED where patients were the only "customers", versus that of the same group of EPs in a CAED where patients and residents generate competing demands for EPs’s time. First, we quantified the time EPs reallocate to resident supervision at the CAED. Second, we determined the categories of activities from which EPs shaved time in order to accommodate supervision.
3.2. Methods

3.2.1. Study Site

We conducted an observational, time-motion study at a 25-bed CAED versus a 15-bed community ED. Both are Level II Trauma Centers within a four-hospital health system in northern Illinois. The CAED is affiliated with two Accreditation Council for Graduate Medical Education (ACGME) accredited residency programs. The community ED does not have ACGME trainees. Table 3.2 shows key performance metrics for both EDs during the fiscal year of 2017.

The EDs have comparable attending physician staffing ratios despite a 10,000 patient census differential between the CAED and community ED. The same group of EPs staff both EDs in our study. Both EDs have double EP coverage from 9am to midnight and single coverage overnight. During the double-coverage hours, each EP is supported by a separate team of nurses and is responsible for an evenly-divided subset of rooms. Chart documentation is completed via dictation or by typing directly into the electronic health records (HER). Neither ED has scribes. This study was approved by institutional IRBs.

Most residents in our study were second (PGY-2) and third (PGY-3) year EM residents. All patients at the CAED evaluated by a resident are staffed by an EP as well, which corresponds to practice previously described (Cobb et al., 2013). In general, a resident takes the patient’s history, performs a physical examination, and then presents the findings to the EP. During the presentation, the resident ideally proposes a differential diagnosis and management plan. The EP then provides feedback to the resident
Table 3.1. Performance Metrics for community-academic and community in FY2017

<table>
<thead>
<tr>
<th></th>
<th>community-academic</th>
<th>community</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED Visits</td>
<td>37,042</td>
<td>27,555</td>
</tr>
<tr>
<td>% Observation Admissions</td>
<td>13.3%</td>
<td>14.5%</td>
</tr>
<tr>
<td>% Inpatient Admissions</td>
<td>14.7%</td>
<td>13.5%</td>
</tr>
<tr>
<td>LWBS</td>
<td>708</td>
<td>352</td>
</tr>
<tr>
<td>Pediatric Volume</td>
<td>3993</td>
<td>4881</td>
</tr>
<tr>
<td>Median Admission Turnaround Time</td>
<td>271</td>
<td>242</td>
</tr>
<tr>
<td>Median Discharge Turnaround Time</td>
<td>196</td>
<td>178</td>
</tr>
<tr>
<td>Pct. Female</td>
<td>55.3%</td>
<td>56.0%</td>
</tr>
<tr>
<td>Pct. Male</td>
<td>44.7%</td>
<td>44.0%</td>
</tr>
<tr>
<td>ESI: level 1</td>
<td>1.0%</td>
<td>0.3%</td>
</tr>
<tr>
<td>ESI: level 2</td>
<td>28.1%</td>
<td>20.0%</td>
</tr>
<tr>
<td>ESI: level 3</td>
<td>52.5%</td>
<td>56.4%</td>
</tr>
<tr>
<td>ESI: level 4</td>
<td>16.3%</td>
<td>20.3%</td>
</tr>
<tr>
<td>ESI: level 5</td>
<td>0.4%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Medicine Admissions</td>
<td>8609</td>
<td>6363</td>
</tr>
<tr>
<td>Surgery Admissions</td>
<td>1831</td>
<td>1420</td>
</tr>
<tr>
<td>Avg. Hourly Patients Seen per Shift</td>
<td>2.8</td>
<td>2.2</td>
</tr>
</tbody>
</table>

(i.e. probing and clarifying questions, nodding, correcting, and using case vignettes from other related cases as supporting evidence). In this study, we consider such care-related, direct EP-resident interaction as the main activity of supervising residents. After the discussion, the EP examines the patient independently with or without the
resident. After the independent evaluation, the EP continues supervision by evaluating the resident’s assessment and modifying the treatment plan together with the resident, if needed. Orders for medications, procedures, and laboratory tests are entered into the EHR and communicated to the nurse, either by the resident or by the EP herself. Residents and EPs both write notes on the patient. The resident notes are generally complete history and physical examination findings, as well as assessments and plans that comprise the medical decision making. EPs generally write an abbreviated supervisory or Physicians at Teaching Hospitals (PATH) note as stipulated by CMS. Importantly, the residents at the CAED do not take the EPs role in all these care-related activities. The EP remains the chief provider of care, with the supervision of residents being an added responsibility.

3.2.2. Data Collection

We followed the activity categorization developed by Tipping et al. (2010a,b) in which they suggested to include the primary categories of direct and indirect patient care. We also followed their definition of direct patient care as those activities involving face-to-face interaction between the [EP] and the patient (Tipping et al. 2010a,b). In our study, all the time that EPs spent in patient rooms is quantified as direct patient care. Accordingly, bedside charting or teaching is counted towards direct patient care because the EPs are performing these activities in the presence of patients. Indirect patient care includes activities relevant to the patient’s care but not performed in the presence of the patient (Tipping et al. 2010a). Following Tipping et al. (2010a) we adjusted to the specifics of our sites by adding customized subcategories. We kept refining
the sub-categories to ensure that they were easily observable and identifiable without subjective interpretation from the observer during a pilot study (Tipping et al. 2010a). Two experienced EPs in our research team helped finalize the categorization (See Table 3.2). The multi-tasking activity only involved an EP simultaneously communicating with other non-resident providers while working on EHR. A resident presenting a case while the EP is reviewing the related EHR is categorized as supervision.

3.2.3. Hypothesis Development

3.2.3.1. Hypothesis 1. A recent study by Hexom et al. reports a mean supervision time by EM faculty on residents of 60.8 minutes over 8-hour observation periods (Hexom et al. 2017). Chisholm et al. also demonstrated that overall EM faculty devoted 11.9% of their time to resident supervision (Chisholm et al. 2004). Therefore we expected to see EPs in our study tailoring resident supervision to their workflow at the CAED.

3.2.3.2. Hypothesis 2. In order to fulfill the responsibility of supervising, EPs have to reallocate their time spent on other care-related activities. We hypothesized that EPs may delegate portions of indirect patient care activities to residents, such as communication and EHR work.

3.2.3.3. Hypothesis 3. It was unclear how the time EPs spend on direct patient care would differ between ED settings. On one hand, patient care by residents might reduce the time that EPs spend on direct patient care. On the other hand, bedside teaching of residents would increase this time. It was also unclear which of these effects would materialize in practice. EPs may constrain the autonomy of residents for reasons of efficiency and safety. Previous studies show limited use of bed-side teaching in practice
Table 3.2. Performance Metrics for community-academic and community in FY2017

<table>
<thead>
<tr>
<th>Primary Category</th>
<th>Main Category</th>
<th>Sub-category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Patient Care</td>
<td>Direct Patient Care</td>
<td></td>
</tr>
<tr>
<td>Indirect Patient Care</td>
<td>Electronic Health Record (EHR)</td>
<td>Reviewing Charting and putting orders</td>
</tr>
<tr>
<td></td>
<td>Communication</td>
<td>Phone calls and consults</td>
</tr>
<tr>
<td></td>
<td>Multi-tasking</td>
<td>Face-to-face communication with other providers</td>
</tr>
<tr>
<td>Supervision</td>
<td>Supervising residents</td>
<td>Communicating with other non-resident providers while working on EHR</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>Personal Transit/travel</td>
</tr>
</tbody>
</table>

(Chisholm et al. 2004; Aldeen and Gisondi 2006). We hypothesized that time spent by EPs on direct patient care at the CAED would be less than that at the community ED.
3.2.4. Pilot Observation

In October 2016, we conducted nine pilot observations totaling 1,764 minutes to design the data collection process and to train the observer (a third-year Ph.D. candidate in Operations). The observer was trained to be responsive to EP’s activity transition and to rapidly track them down. During the pilot study, the observed EPs explained to the observer his/her current activity (e.g. "I am charting” or "I am going to have lunch"). The observer also asked the EPs about ambiguous situations to avoid misinterpreting the activities being performed. Because we only had a single observer in the study, there was no concern for inter-observer reliability or consistency in measurement.

3.2.5. Selection of Participants

We selected a nonrandomized convenience sample from a total of 51 EPs in this health system. The primary inclusion criteria required the participating EP to work at both EDs during the study. Authors were excluded from the sample. This resulted in 30 eligible EPs, from which 25 EPs verbally consented to participation. They were informed that the study was about ED workflow but were blinded from the specific objective. Data confidentiality was ensured.

3.2.6. Study Protocol

We observed each EP twice at each ED for a total of 100 observations (50 at the CAED and 50 at the community ED.) The length of a single observation session was 240 minutes (Fuchtbauer et al. 2013; Chisholm et al. 2011), totaling 400 observation hours. Importantly, the EPs served as their own control. The observed shifts were a convenience
sample, but were selected to be approximately evenly-distributed over the days of the week and AM/PM shifts.

Residents, nurses, and advanced practice providers were not observed directly, but their interactions with the observed EP were recorded. The observer shadowed an individual EP continuously except when the EP was inside the patient room or requested privacy. Tracked information includes locations, categorized activity, with start and end times for activity executions. All observational data were recorded using an iPad app called Eternity. To minimize the Hawthorne effect, the observer maintained a safe distance from the observed EP and did not engage in conversation during the observation periods. To protect patient privacy, the observer did not enter patient rooms. The iPad was held by the observer and was obscured from the EP during observations.

We calculated Patient load and Leave without being seen (LWBS) rate for each observation using the patient-level turnaround data derived from Data Warehouse. Patient load was calculated by adding the number of new patients assigned to the observed EP and received care during the observation, to the number of patients already under the observed EP’s care at the beginning of the observation (Tipping et al. 2010a). We also calculated the percentages of patient load in each Emergency Severity Index (ESI) level. ESI provides clinically relevant stratification of patients into five levels from 1 (most urgent) to 5 (least urgent) based on acuity and resource needs (Gilboy N, Tanabe T, Travers D). The LWBS rate was defined as the ratio of the number of LWBS new patients (i.e. new patients who arrived during the observation and were assigned to the observed EP, but subsequently left without being seen by this EP) (Polevoi et al.)
to the number of total new patients who arrived during the observation and were assigned to the observed EP.

The supervised residents were queried confidentially using a survey compiled from previously reported surveys (Holt et al. 2010). Residents were asked to rate "How sufficient is the supervision you received from attending in the past four hours on a three-point Likert scale with the descriptors: Very sufficient, Sufficient, and Not at all/slightly sufficient (Jacoby and Matell 1971)." The residents were also asked to "Describe your learning outcome in the past four hours using a three-point scale with the descriptors: I have learned a significant amount, I have learned something, and I didn’t learn anything". These surveys were completed shortly after the observations (within 2 hours).

3.2.7. Measurements and Primary Data Analysis

The time spent on each activity was measured in the unit of second. Mean time for each main category at community ED vs. CAED were presented in percentages (out of a 240-minute observation). We then reported the detailed time spent by EPs on each sub-category across two EDs during observations. We also reported the time an EP spent per patient load on each sub-category during an observation, using the total time the EP spent on the sub-category divided by the patient load during this observation (Tipping et al. 2010a).

The results of Shapiro-Wilk test suggest that most of our time-spent data did not follow normal distribution (Table 3.3 and 3.4); we thereby presented the median and interquartile ranges (IQR) of minutes spent on each sub-category at CAED vs. community ED. A Wilcoxon two-sample test was performed. To control the Type I error, we
Table 3.3. Shapiro-Wilk Test for Normality of Total Minutes Spent on Sub-category at community-academic vs. community

<table>
<thead>
<tr>
<th>Sub-category</th>
<th>community-academic: p-value</th>
<th>community: p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct patient care</td>
<td>0.024*</td>
<td>0.008*</td>
</tr>
<tr>
<td>EHR: reviewing</td>
<td>0.105</td>
<td>0.012*</td>
</tr>
<tr>
<td>EHR: charting and putting orders</td>
<td>0.419</td>
<td>0.349</td>
</tr>
<tr>
<td>Communication: Phone calls and consults</td>
<td>0.012*</td>
<td>0.011*</td>
</tr>
<tr>
<td>Communication: Face-to-face with other providers</td>
<td>0.002*</td>
<td>0.069</td>
</tr>
<tr>
<td>Multi-tasking</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Supervising residents</td>
<td>0.218</td>
<td>-</td>
</tr>
<tr>
<td>Personal (e.g. eating, restroom, socialization)</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Travel</td>
<td>0.014*</td>
<td>&lt;0.001*</td>
</tr>
</tbody>
</table>

* For the significance level of 0.05, an observed time spent data with a p-value less than 0.05 rejects the null hypothesis that the data are from a normally distributed population.

corrected the p-values using the Benjamini-Hochberg procedure with a false discovery rate (FDR) of 0.05 (Benjamini and Hochberg 1995; Genovese et al. 2002). Two-sided FDR-adjusted p-values < 0.05 are considered statistically significant. All analyses were run on R version 3.4.3 (Team 2018).

3.3. Results

During the formal observational study from March 2017 to August 2017, the 25 observed EPs performed a total of 35,348 executions of the subcategorized activities
Table 3.4. Shapiro-Wilk Test for Normality of Average Minutes Spent per Patient on Sub-category at community-academic vs. community

<table>
<thead>
<tr>
<th>Sub-category</th>
<th>community-academic:</th>
<th>community:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value</td>
<td>p-value</td>
</tr>
<tr>
<td>Direct patient care</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>EHR: reviewing</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>EHR: charting and putting orders</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Communication: Phone calls and consults</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Communication: Face-to-face with other providers</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Multi-tasking</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Supervising residents</td>
<td>&lt;0.001*</td>
<td>-</td>
</tr>
<tr>
<td>Personal (e.g. eating, restroom, socialization)</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Travel</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
</tbody>
</table>

* For the significance level of 0.05, an observed time spent data with a p-value less than 0.05 rejects the null hypothesis that the data are from a normally distributed population.

during the entire 400 observation hours. On average 355 activities were performed per observation at the CAED vs. 332 at the community ED (P = 0.11). The average number of patient load across observations was 15 at the CAED vs. 13 at the community (P = 0.40). The ESI scores of these patients were significantly different between the CAED vs. community for levels 1-3: Level 1 (1.1% vs. 0.1%, P = 0.008); Level 2 (37.7% vs. 20.7%, P < 0.001); Level 3 (47.9% vs. 62.7%, P < 0.001). There was no difference in the average number of patients discharged by the participating EP during
the observations (CAED=7 vs. community=7, P=0.59). The average LWBS rates were similar (CAED = 1% vs. community = 0.9%, P = 0.33).

Figure 3.1 illustrates the time utilization profile of an average EP per 240-minute observation session. EPs spent 34.2 minutes (14.2%) supervising residents (8.5 minutes per hour). Direct patient care accounted for 76.8 minutes (32%) vs. 79.9 minutes (33.3%) (P = 0.31) at the CAED and community ED respectively. Indirect patient care accounted for 99.8 minutes (41.6%) and 128.4 minutes (53.5%) at the CAED and community ED respectively (P < 0.001). Non care-related activities (personal time and travel) accounted for 12.1% vs. 13.1% (P = 0.96) respectively and did not statistically differ.

Comparing the CAED to the community ED, significant median time decreases were found in EHR review (32.2 minutes vs. 23.9 minutes, FDR-adjusted P = 0.003),
charting and putting orders (41.6 minutes vs. 36.7 minutes, FDR-adjusted P = 0.029), face-to-face communication (25.8 minutes vs. 17.9 minutes, FDR-adjusted P = 0.002), phone calls/consult communication (17.1 minutes vs. 8.4 minutes, FDR-adjusted P < 0.001), and multi-tasking (7.2 minutes vs. 4.32 minutes, FDR-adjusted P = 0.031). Personal time did not change significantly.

Median minutes spent per patient load on communication, either face-to-face or by phone, at the CAED decreased by almost 40%. EPs working at the CAED spent 1.14 median minutes less reviewing in EHR (FDR-adjusted P = 0.028). No significant change was found with respect to direct patient care, charting and putting orders in EHR, multi-tasking and other non-patient care activities. Survey results Thirty-one residents completed 47 unique session-survey responses (response rate is 78.3% from 60 session surveys). One hundred percent of the responses described the supervision as "very sufficient." Forty-three percent of the responses reported having "learned something" and 57% reported having "learned a lot" during the corresponding sessions.

3.4. Discussion

Building on DeBehnke’s call for a more refined understanding of the time and effort expended on educating and supervising residents, we studied how EPs adjust their clinical practices when resident supervision is added to their responsibilities. The primary strength of this study is an extensive dataset and a subject group that served as their own controls. To our knowledge, ours is the first time-motion study to fully map the time utilization profile of EPs working with and without residents. While this
information represents two EDs and mostly upper level residents, we believe it is a reasonable starting point for other studies on this important topic.

The key findings are: 1) EPs spent a substantial portion of their clinical time supervising residents; 2) EPs spent the majority of their clinical time in direct patient care; and 3) EPs experienced a significant reduction in indirect patient care when working with residents.

First, our EP cohort spent 14.2% of their time supervising residents. This translates to 68 minutes over an 8-hour shift, consistent with the supervision time found by Hexom et al. (2017). Time spent interacting with EPs may determine perceived education quality by the residents. This also echoes the positive responses from the residents to our survey.

Second, we confirmed that our EPs spent the majority of their time performing direct patient care (Hexom et al., 2017; Fuchtbauer et al., 2013; Chisholm et al., 2004). More importantly, even though the EPs devoted a substantial amount of their time to resident supervision at CAED, direct patient care time did not change significantly. Direct patient care, the primary priority for our EPs, were preserved despite substantial effort towards supervision. Consistent with the findings by Chisholm et al. (2004), seldom do the EP and the resident at the CAED work simultaneously at the patient’s bedside, except for verification of resident findings on history and physical or procedural supervision.

Third, we observed that time savings from EPs’ offloading indirect patient care activities to the residents largely offset the supervision time: EM residents contributed to indirect patient care and expanded EPs capacity in addition to the direct patient care
they provided a ‘win-win situation’ of resident supervision. In our setting, adding an intermediate or upper level EM resident to the CAED team extended the EPs’ ability to evaluate approximately 10,000 more patients per year (who were arguably sicker by our ESI score) with comparable staffing and the same bedside time.

Finally, residents freed EPs from specific indirect patient care activities such as communication and EHR work. At both aggregate and per patient-load levels, EPs delegated most of the communication to their residents, spending significantly less time making phone calls or face-to-face communication to other healthcare providers. The EPs providing supervision in exchange for release from the tasks can be viewed as an apprenticeship-type of experience: "The resident will perform these indirect patient care tasks under my supervision, and in return, I will provide the resident with diagnostic feedback on how to perform them better and more effectively." The presence of residents also endows the EPs with discretion as to how and when to expense their efforts a ‘currency of resident apprenticeship’. Several of our EPs suggested that when working at the CAED, they delegated more of the charting to residents when the patient load was low, but would complete more of it on their own when the patient load was high so that residents can help do more direct patient care to accelerate the patient throughput. When working at the community ED, their time spent per patient on charting is relatively stable and independent of the patient load. This anecdotal evidence is consistent with the substantially wider IQR of per-patient load EHR charting time at the CAED (Table 3.5 in Supplement).
3.5. Limitations and Future Work

Our results should be interpreted in the light of several limitations. First, some possible confounding factors were beyond our control. For example, the CAED had a larger concentration of high acuity patients both anecdotally and suggested by their ESI scores. This would bias upwards the observed change in time on direct patient care at CAED, as critical patients typically require more direct patient care. Such bias was difficult to avoid without randomizing patients across EDs and EPs. Meanwhile, ESI has been criticized for its low accuracy and high variability in clinical practice (Mistry et al. 2018; Buschhorn et al. 2013). Observing differences in patient ESI distributions thereby may not suffice to confirm the difference in the actual acuity distributions across two EDs. Patient load was another potential confounder. With high patient load, EPs probably deferred documentation to make more time for direct patient care. We tried to balance the distribution of patient loads across two EDs by spreading the observations over days of week and AM/PM shifts. The average number of patient load across observations at CAED turned out to be higher than at the community ED (15 vs. 13). This difference was not statistically significant ($P = 0.40$), probably because of the large variance in patient loads, but anecdotally, the EPs "feel the difference of simultaneously carrying more patients at the CAED". To further adjust for the patient load, we reported the average time spent per patient load on each sub-category in Table S.4. Future study with more observations may achieve more consistent estimate and comparison of patient loads.

A potentially stronger method to eliminate these biases introduced by systematical differences in different EDs is to compare the same EP at the same ED on shifts
with versus without residents. Although we did not find such opportunity, Salazar et al. (2001) achieved this by using a resident strike period as the control, and compared quality indicators of patient care during days when residents were on duty versus on strike.

Second, we did not separate the time spent on bedside teaching from direct patient care time. This observation rubric followed prior reported studies (Tipping et al. 2010a). Furthermore, the supervision time we recorded already accounted for a major part of the total EM resident-faculty interaction time as shown in previous study (Chisholm et al. 2004).

Third, we did not capture EPs’ time spent after shifts. EPs might have to stay late after shifts to complete patient notes. More data capturing EPs’ work after shifts would provide a more holistic analysis of resident effect on their time utilization profile.

Fourth, the accuracy of the single non-clinical observer’s interpretation, as well as the Hawthorne effect may limit the results. But we believe that the pilot observations totaling almost 30 hours provided sufficient practice for the observer to achieve reasonable recording accuracy and to remain as unobtrusive as possible during observations (Tipping et al. 2010a). Besides, the inherent bias of the observer’s assessment would be carried across both EDs and would not affect the comparison results. Having multiple observers may make the study more robust and replicable, but that imposes a higher requirement for resources and study design (Lopetegui et al. 2013; Zheng et al. 2011).

Fifth, we did not capture overnight shifts in our observations, concerning that the difference in coverage and availability of other medical providers may confound the
results. Future research specifically focusing on potential differences in the resident effects on overnight shifts are warranted.

Sixth, these results may not be generalizable to traditional university-based teaching hospitals and other CAEDs, where staffing, supervision, and institutional culture can be quite different. For example, our resident sample consisted entirely of intermediate and upper level residents. The results may not generalize to institutions that were staffed primarily with junior trainees (i.e. medical students and interns), as senior residents are reported to be more productive than junior trainees (Jeanmonod et al. 2008; Lammers et al. 2003; Brennan et al. 2007).

Finally, we had limited number of observations per EP per ED. A larger sample size, with a moderate-large number of observations per EP, per ED, or even under different levels of patient load and acuity, would support a more detailed empirical study that goes beyond comparison of descriptive statistics aggregated at ED level.

3.6. Conclusion

EPs working with EM residents in a community-academic ED reap significant time savings from the responsibilities of indirect patient care, and remunerate those savings in kind to the residents in the form of supervision which accounted for 14.2% of their clinical time. These time savings allow them to foster a clinical learning environment where residents and fellows can interact with patients under the guidance and supervision of qualified faculty members who give value, context, and meaning to those interactions and achieve "balance of service and education" in residency training (for
More importantly, it allows the emergency physician to preserve their ability to provide direct high quality safe patient care, which remains their core mission.

3.7. Appendix of Chapter 3
Table 3.5. Average Minutes per Patient (during the 240-minute Observation) Spent on Sub-category at community-academic vs. community

<table>
<thead>
<tr>
<th>Sub-category</th>
<th>community-academic: Median (IQR)</th>
<th>community: Median (IQR)</th>
<th>Raw P-Value</th>
<th>FDR-adjusted P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct patient care</td>
<td>5.46 = 5min28sec (3.45, 11.24)</td>
<td>5.97 = 5min58sec (4.72, 9.81)</td>
<td>0.43</td>
<td>0.553</td>
</tr>
<tr>
<td>EHR: reviewing</td>
<td>1.58 = 1min35sec (1.08, 3.44)</td>
<td>2.72 = 2min43sec (2.09, 3.95)</td>
<td>0.01*</td>
<td>0.028*</td>
</tr>
<tr>
<td>EHR: charting and putting orders</td>
<td>2.97 = 2min58sec (1.73, 4.88)</td>
<td>3.39 = 3min24sec (2.56, 5.03)</td>
<td>0.22</td>
<td>0.33</td>
</tr>
<tr>
<td>Communication: Phone calls and consults</td>
<td>0.62 = 37sec (0.21, 0.98)</td>
<td>1.36 = 1min22sec (1.01, 2.06)</td>
<td>&lt; 0.001***</td>
<td>&lt; 0.001***</td>
</tr>
<tr>
<td>Communication: Face-to-face with other providers</td>
<td>1.36 = 1min22sec (0.73, 3.27)</td>
<td>2.15 = 2min9sec (1.37, 3.15)</td>
<td>&lt; 0.01**</td>
<td>0.042*</td>
</tr>
<tr>
<td>Multi-tasking</td>
<td>0.35 = 21sec (0.11, 1.04)</td>
<td>0.58 = 35sec (0.41, 0.87)</td>
<td>0.09</td>
<td>0.156</td>
</tr>
<tr>
<td>Supervising residents</td>
<td>2.22 = 2min13sec (1.58, 4.58)</td>
<td>0</td>
<td>&lt; 0.001***</td>
<td>&lt; 0.001***</td>
</tr>
<tr>
<td>Personal (e.g. eating, restroom, socialization)</td>
<td>1.54 = 1min32sec (0.52, 2.47)</td>
<td>1.45 = 1min27sec (0.62, 2.76)</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Travel</td>
<td>0.63 = 38sec (0.46, 0.99)</td>
<td>0.64 = 38sec (0.45, 0.87)</td>
<td>0.46</td>
<td>0.517</td>
</tr>
</tbody>
</table>

Note: * P<0.05; ** P<0.01; *** P<0.001.
References


Laila Maria Fuchtbauer, Birgitte Nrgaard, and Christian Backer Mogensen. Emergency department physicians spend only 25% of their working time on direct patient care. *Danish medical journal*, 60(1):A4558, jan 2013. ISSN 2245-1919.


Raghav Pavan Karumur, Bowen Yu, Haiyi Zhu, and Joseph A Konstan. Content is king, leadership lags: Effects of prior experience on newcomer retention and productivity in online production groups. 2018.


Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. The emerging role of data scientists on software development teams. In *Proceedings of...*


José M Montoya and Dave Raffaelli. Climate change, biotic interactions and ecosystem services. 2010.


Ikuiro Nonaka. A dynamic theory of organizational knowledge creation. 


[https://www.R-project.org/](https://www.R-project.org/)


