Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Drivers of Knowledge Base Adoption, Analysis of Czech Corporate Environment

Author: **Bc. Zuzana Rakovská**

Supervisor: **PhDr. Václav Korbel**

Academic Year: **2014/2015**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, July 15, 2015

Signature

Acknowledgments

The author would like to express sincere thanks to company SEMANTA, s.r.o. which contributed importantly to this thesis not only by providing access to their systems' databases but also by maintaining the background essential for hypotheses' creation. Special thanks belongs to Mgr. David Voňka for his help in theory and practice, and his attentiveness and eagerness in finding solutions in every aspect.

The author is also heartily grateful to PhDr. Václav Korbel for his factual advice, valuable suggestions and professional approach. Lastly, the thanks is offered to Ing. Katarína Rakovská and Kamil Gric for their supportive and cheerful attitude which accompanied the author during writing this thesis.

Abstract

This thesis analyses the process of knowledge-base adoption in the enterprise environment. Using data from two knowledge-management systems operated by the company, Semanta, s.r.o. we studied the day-to-day interactions of employees using the system and identified the important drivers of system adoption. We began by studying the effect of co-workers' collaborative activities on knowledge creation within the system. It was found that they had a positive and significant impact upon overall knowledge creation and thus on adoption. Secondly, we explored how the newly defined concept of gamification could help determine and encourage an increase in knowledge creation. The use of gamification tools, such as the "Hall of Fame" page, turned out to have significant influence in the adoption process. Thirdly, we examined how users continually seek knowledge within the system and how asking for missing information and being supplied with answers has an impact on adoption rates. It was shown that the quicker the responses and the more experts dealing with requests the greater the impact on knowledge base adoption. Finally, we showed that the size and character of the company deploying the knowledge management system does not influence the adoption drivers. This thesis represents an effort to fill the literature gap surrounding effective knowledge-base adoption in an intra-company environment. Moreover, as far as we know, it represents the first attempt to estimate the relationship between gamification concepts and knowledge-base adoption not only in the Czech Republic but also worldwide.

JEL Classification	J24, O15, O34, O35, O52,
Keywords	knowledge base, gamification, knowledge-base adoption, knowledge management system, technology adoption, intellectual capital

Author's e-mail	zuzana.rak@gmail.com
Supervisor's e-mail	vaskor@email.cz

Abstrakt

Tato diplomová práce analyzuje proces přijetí znalostní báze v podnikatelském prostředí. Použitím dat z dvou systémů pro znalostní management provozovaných společností Semanta, s.r.o. jsme studovali každodenní interakce mezi

zaměstnanci jako uživateli systému a identifikovali jsme důležité faktory působící na přijetí tohoto systému. Za prvé jsme v práci studovali, jak společná aktivita pracovníků ovlivňuje tvorbu znalosti v systému. Našli jsme významný a pozitivní vliv tohoto faktoru na celkovou tvorbu znalostí, a tedy i na přijetí systému znalostního managementu jako takového. Za druhé jsme zkoumali, jak nově definovaný koncept "gamifikace" může podpořit zvýšení tvorby znalosti. Výsledky regrese ukázaly, že používání "gamifikovaných" nástrojů, jakým je například stránka "Hall of Fame", má významný vliv na proces přijetí znalostní báze. Za třetí jsme studovali, jak uživatelé kontinuálně hledají znalost v systému a jaký účinek na přijetí znalostní báze má požadování chybějící informace a následné získávání těchto odpovědí. Ukázali jsme, že rychlejší odpovědi a větší počet expertů, kteří se otázkami zabývají, pozitivně působí na přijetí. A konečně studie také prokázala, že velikost a charakter společnosti, která systém znalostního managementu zavádí, nemá vliv na faktory přijetí. Tato diplomová práce představuje snahu o vytvoření chybějící literatury, která studuje efektivní přijetí znalostních bází ve firemním prostředí. Pokud je nám známo, představuje tato práce první pokus prokázat vztah mezi konceptem "gamifikace" a přijetím znalostní báze nejenom v České republice, ale také celosvětově.

Klasifikace JEL	J24, O15, O34, O35, O52,
Klíčová slova	znalostní báze, gamification, přijetí znalostní báze, systém znalostního managementu, přijetí technologie, intelektuální kapitál

E-mail autora	zuzana.rak@gmail.com
E-mail vedoucího práce	vaskor@email.cz

Contents

List of Tables

List of Figures

Acronyms

KB Knowledge Base

KMS Knowledge Management System

Guides Encyclopaedia Guides Knowledge Base

MO The Mobile Operator's Knowledge Base

Master's Thesis Proposal

Author	Bc. Zuzana Rakovská
Supervisor	PhDr. Václav Korbel
Proposed topic	Drivers of Knowledge Base Adoption, Analysis of Czech Corporate Environment

Motivation Knowledge - an intellectual capital asset - is considered as the basic economic resource that is fundamentally embedded in the workers who perform the job-specific tasks. It bears all the features of the resources´ definition, and thus, provides high competitive value to an organization. An effective accumulation, preservation and sharing of knowledge within a company - knowledge management (KM) - is a critical success factor in a fast changing business environment. A KM deployment is not limited to installation of knowledge base technology. Its cornerstone is a process of user adoption and innovation-affected cultural change in human behavior. Therefore, the proper assessment of factors affecting the process of KM adoption seems as a powerful tool in obtaining a competitive advantage. Only a few studies have analyzed factors affecting behavioral intentions (BI) to use innovations so far. Davis (1989) introduced Technology Acceptance Model (TAM) based on users perceptions in order to explain this phenomenon. The model has been further extended in several studies (Li-Su Huang, 2014; Suresh, 2013; Ren-Zong Kuo, 2011). Ren-Zong Kuo (2011) indicated that the effective KB accumulation is not feasible without users´ willingness to share their knowledge. People may not intend to share their unique knowledge due to a fear of losing their power position in the organization. Therefore he suggests promotion of a knowledge-sharing culture, such as reward system, reputation etc. The solution to such an enhancement can be found in the new concept, called gamification (Leeson, 2013). Gamification can be defined as the use of game mechanics and experience design to digitally engage and motivate people to achieve their goals (gartner.com).

Thus, such mechanics are able to instrument KM adoption and via its processes can help explain studied factors. Moreover gamification itself behaves as a certain kind of nudge that places behavioral aspects into the knowledge economics. The thesis will be written in cooperation with SEMANTA, s.r.o. - a company that develops and implement knowledge base for enterprise clients. Its platform comprises gamified tools, such as comments, likes and shares. Therefore it provides the unique dataset that can be used for further analysis of the stimuli of KB adoption exerting not only behavioral intentions themselves but also nudges included in the principle of gamification. Such driving forces are then the powerful tools for the identification of effectiveness of investment into different types of activities that support widespread intracompany adoption of KB usage.

Hypotheses

1. Further content creation (creation of pages, comments, page edits, etc) depends on ?gamified tools? such as thanks for previously edited/created pages, comments or hall of fame placement, etc.

2. Fast response to comments drives the KB adoption.

3. Integration of KB with other platforms enhances usage of these platforms.

Methodology The unique dataset that will be used for the analysis is provided by company SEMANTA, s.r.o.. It consists from numerous observations that are of the "big-brother" character. It means that the dataset catches all the platform-users activities in the certain time horizon. In order to obtain variables the intensive extraction and transformation of the dataset would be needed.

The hypotheses stated above are expected to be tested employing the following variables:
Hypothesis #1:

- Dependent variable (alternatives):
  - Number of pages created in a week
  - Number of comments added
  - Number of page edits
- Independent variables:
  - Demographics/individual

  - Number of visits to previously edited/created pages
  - Thanks for previously edited/created pages
  - Comments to my pages
  - Hall of fame placement (where: homepage? Monitoring centre?)
  - KB size
  - KB rate of growth

Hypothesis #2:

- Dependent variable:
  - Dummy (=1 if a user comes back to ency within a sufficiently short period after asking a question)
- Independent variables:
  - Time to answer of the question (=30min if the question was answered 30 minutes after it was asked)
  - Demographics of the asking user
  - Identity of the answering person
  - Area of the question (very technical, business?)

Hypothesis #3:

- Dependent variable:
  - Number of visits of individual reports
- Independent variables:
  - Reporting platform integration in place (Yes/No)
  - Report complexity
  - Number of KB definitions relevant for this report
  - KB-unretated determinants of report popularity

Based on the longitudinal structure of the dataset I will estimate the model using panel data methods.

Outline

1. Motivation
2. Literature Review & Theoretical Background
3. Data - extraction and transformation of the dataset
4. Methodology
5. Results
6. Discussion
7. Conclusion

Expected Contribution I will conduct the assessment and description of factors affecting the KM adoption. These drivers will be analyzed following the

unique dataset that catches all the platform-users activities in the certain time horizon. Moreover, it includes processes of newly defined concept of gamification (f.e. record of shares, comments, etc.) that can be used as the tool enforcing the KM adoption (Leeson, 2013). Such an analysis has not been established for the Czech business environment yet. Since the topic concerning the KM adoption is relatively new, this thesis would be also one of the first papers studying this phenomena.

Core bibliography

1. Braganza, A., R. Hackney & S. Tanudjojo (2009): "Organizational knowledge transfer through creation, mobilization and diffusion: a case analysis of InTouch within Schlumberger." *Information Systems Journal* **19(5)**: pp. 499–522.
2. Davis, F.D. (1989): "Perceived usefulness, perceived ease of use, and user acceptance." *MIS Quarterly*, **13(3)**: pp. 319–340.
3. Hamari, J., J. Koivisto & H. Sarsa (2014): "Does Gamification Work? - A Literature Review of Empirical Studies on Gamification." In *proceedings of the 47th Hawaii International Conference on System Sciences,* Hawaii, USA.
4. Huang, Li-Su., & Cheng-Po. Lai (2014): "Knowledge Management Adoption and Diffusion Using Structural Equation Modeling." *Global Journal of Business Research (GJBR).* **8(1)**: pp. 39–56.
5. Kuo, Ren-Zong. & Gwo-Guang. Lee (2011): "Knowledge management system adoption: exploring the effects of empowering leadership, task-technology fit and compatibility." *Behaviour & Information Technology* **30(1)**: pp. 113–129.
6. Leeson, T. C. (2013): "Driving KM behaviors and adoption through gamification." *KM World*[online] **22(4)**: pp. 10–12.
7. Massa, S. & S. Testa (2007): "ICTs adoption and knowledge management: the case of an e-procurement system." *Knowledge & Process Management* **14(1)**: pp. 26–36.
8. Ritchie, W.J., S.A. Drew, M. Sprite, P. Andrews & J.E. Carter (2011): "Application of a Learning Management System for Knowledge Management: Adoption and Cross-cultural Factors." *Knowledge & Process Management* **18(2)**: pp. 75–84.
9. Rogers, E.M. (1995): *The diffusion of innovation.* 3rd ed. New York: Free Press.
10. Suresh, A. (2013): *"Knowledge Management Adoption, Practice and Innovation in the Indian Organizational Set Up: An Empirical Study." Journal of Information Technology & Economic Development.* **4(2)**: pp. 31–42.

Author Supervisor

Chapter 1

Introduction

The world as of today depends highly on exchange of information, its processing and utilization. Knowledge represents new intangible asset that companies accumulate and use to achieve their business goals. Effective knowledge management is capable of inducing cost reduction as well as creating competitive advantage in the market. However, the extraction of such benefits does not depend on installation of knowledge management system. Its cornerstone is knowledge-base adoption by firms' culture.

Every employee, not only directors and managers, possesses a certain knowledge that is unique for company. For example, one might know how to provide best services to customer, another one is experienced in product design and there might be a project manager who knows how to lead a project to be profitable. All these workers represent company's intellectual capital that is essential in creating competitive products and services. However, if such acquired knowledge remains only in their minds, company might simply lose part of its know-how when the employee leaves. To prevent this, many firms are deploying knowledge management systems (KMS). It is a widely spread solution that captures workers' unique insights and stores them into knowledge base. As a result, such collected experience is transformed into corporate one which can not be simply removed because firm is now able to control it. Moreover, all knowledge is stored in one place that is available to every worker.

Knowledge management system can be considered as a "modern production technology" whose output - knowledge, exhibits increasing returns to scale. Firstly, KMS makes its content available to experts who are able to extract any previously used and shared solution, and adapt it to a current problem. It provides expertise to less experienced personnel and also avoid delays when

expertise is needed (Smith 1985). Hence, each such task-execution is facilitated and workers' costs are reduced. For example, a newly hired consultant saves her working hours when using stored knowledge of senior consultants about customers' needs, etc. (Ofek & Sarvary 2001). Secondly, accumulated knowledge creates space for learning from experience and lead to better solutions. Knowledge management system simply keeps record of decisions and actions that are consistent and available over time. This leads to higher-quality processes and services that create competitive advantage and superior performance in the relevant market (Gjurovikj 2000).

In theory, a knowledge management system is a powerful tool in achieving strategic objectives. However, like every production technology, also KMS needs inputs for proper functionality. On the one hand, experts are irreplaceable intake in knowledge-base production (Davenport *et al.* 1989). If they are not willing to share their unique experience to others via knowledge-base channel, benefits from economies of scale are not in place and company is losing competitiveness (Wong & Aspinwall 2005; Ritchie *et al.* 2011). On the other hand, workers that are not sufficiently motivated to seek and use already created knowledge are less effective and increase firm's costs. Hence, incorporation of knowledge management into company's processes is not the final success factor. Employees must be willing to hand over their knowledge and use the corporate information in order to induce cost-efficiency. In other words, knowledge base must be adopted among them.

Although knowledge management has been widely discussed in the last decade, there are only few studies capturing the process of knowledge-base adoption within a firm culture. The respective literature gap results mainly from the subject's novelty and from the lack of empirical data in the intra-company and also in the inter-company level. Knowledge base acceptance is a cornerstone for successful KMS which ensures long term sustainability of its benefits (Huang & Lai 2014; Suresh 2013; Yeoh & Koronios 2010, and others). However, users'/employees' adoption is not self-acting. It needs stimulus through which workers are motivated to create and seek for contents of knowledge base. This thesis hence, represents effort to fill the literature gap on knowledge-base adoption and provides a comprehensive explanation and estimation of drivers affecting knowledge-base adoption.

We center our study on analysis of knowledge bases designed by company Semanta, s.r.o., which develops and deploys knowledge management systems (KMSs) for enterprise clients all over the world. Its KMSs are available through

internet and are in form of web application similar to Wikipedia. It is represented by set of pages organized into trees with hierarchies of classes and subclasses referencing to each other. These pages differ from usual web pages in a sense that every user has access to them and is able to produce their content by editing them, adding new information, creating sub-pages, etc., just like in Wikipedia. In addition, Semanta's KMSs employ additional tool for content creation: inserting comments to already generated pages. Knowledge base is thus a result of collaborative, non-proprietary production process, based on sharing resources and outputs among individuals (Aaltonen & Seiler 2014).

Semanta stores information on every user's action performed in its system. These captured actions are organized in tables in which every row represents detailed information on who did what, when and where (exact page) it happened, etc. We were thus able to extract data capturing history of system-users activity and collaboration with other users or observations on certain actions performed. We have already discussed that there are two parts of knowledge-base adoption considered in this thesis: continuous knowledge (content) creation and continuous knowledge-seeking. Knowledge creation arises when employees generate new pages or when they edit them or comment them. Knowledge seeking means using knowledge base and this is done by visiting its content (pages) by system users. To assess continuous actions we are studying counts of such events (knowledge creation and knowledge seeking) for studied employees in one-week long periods. On the one hand, adoption is induced by factors affecting amount of pages created, edited or commented by an user in consecutive weeks. On the other hand, it is induced by drivers affecting amount of system visits by an user in a week.

Firstly, we analyze activity of other users within knowledge-base space interacting with knowledge creator. Nature of KMSs studied allows users to add small pieces of information relying on subsequent editors or commenters to develop the content further. We consider such collaboration to be strong motivational tool for knowledge creator leading her to generate another content. Hence, the first studied factor is collaborative activity of other co-workers. Secondly, in order to identify other drivers related to content creation, we exploit new concept called gamification. This construct employs those elements from games that engage "players" to stay in game (like points, badges, leader-boards, etc.) and apply them in other non-game contexts (Leeson 2013). Semanta is directly incorporating leader-board-based gamified tool in its KMSs named *Hall of Fame*. It is in a form of page showing users who were the most active in

a previous week and achieved first five positions in different categories. Such tracked category is for example *Commenter* and *Hall of Fame* page shows first five contributors who inserted the highest number of comments into knowledge base in a previous week. We identify the main motivational mechanisms to be: viewing placements reached in *Hall of Fame* leader-board and the incentive resulting from reaching/not reaching the actual placements. Finally, we study drivers of continuous knowledge seeking as the second important part of knowledge-base adoption. We assume that an employee continuously search for precious information in knowledge base when she was satisfied with previous experience in seeking any of it. We employ feature that is a part of Semanta's KMSs that allows workers to ask system's experts to deliver missing knowledge in the base. This is done by using a *Ask* button placed in knowledge bases. Here the analyzed drivers are: the speed with which system experts (users of KMS) answer the request set by other employees and the variety and amount of these answers.

We work with different data in each analyzed hypothesis therefore, our results are estimated using three different methodologies. In first two hypotheses we are dealing with panel data of users across weeks in which dependent variables are weekly amounts of content generated by an employee. Both are suffering from overdispersion, however, in first hypothesis we also detected excess zero problem. As a result, we transformed the first panel into cross-sectional data by using dummy variables to estimate fixed effects and employed zero-inflated negative binomial model (ZINB). The second panel was estimated using random effects negative binomial model (RE NegBin). In third hypothesis we concentrate only on employees that asked experts for a missing knowledge. We study effects on number of visits performed by these users in a week after they obtained answers. Hence, we are not dealing with panel of users across weeks. Instead, the data are structured into cross-section of questions that were once asked by an employee whose activity (further knowledge seeking) is than subject of our analysis. Since overdispersion is present also in this case, we use standard negative binomial model for event counts.

Our framework is innovative in the way that we will assess intra-company interactions between workers as main factors, while the literature concentrates mainly on studying those arising from inter-company relations. This enables us to study direct influences on KMS acceptance on a firm level. Further, as far as we know, this thesis represents the first attempt not only in the Czech Republic but also worldwide, to estimate relation between newly defined con-

cept of gamification and knowledge-base adoption. And finally, employing two knowledge management systems that differ in size of deploying firm allows us to study the importance of knowledge-base size in intra-company environment.

We found overall positive and significant effects of co-workers' collaborative activities on further knowledge creation. Moreover, usage of gamified tools within knowledge bases turned to be another important driver for the content generation. Study of factors affecting knowledge seeking proved that quick responses and number of experts dealing with requests boost knowledge-base adoption. And finally, we showed that the size and character of company deploying knowledge management system does not matter.

The thesis is structured as follows: Chapter 2 summarizes possible approaches in analyzing adoption process and introduce the gamification concept in relation to knowledge management systems. Chapter 3 offers description of knowledge bases designed by Semanta and their elements. In Chapter 4 we discuss studied hypotheses in detail. Chapter 5 characterizes extraction of data, provides its description and defines variables used. Chapter 6 specifies methodology that we work with and Chapter 7 reports results of our empirical research. Chapter 8 summarizes our findings and offers suggestions for further study.

Chapter 2

Literature Review and Theoretical Background

The importance of knowledge management (KM) adoption in corporate environment was emphasized in several studies. Although, this area is very recent, a number of approaches have been developed to examine the forces that impact effective knowledge management implementation. These concepts differ mainly in understanding of knowledge management system (KMS) but also in interpretation of the adoption process. Thus, they can be specified as follows:

1. **Critical Success Factor (CSF) approach** - studies and ranks critical factors that affect successful adoption of knowledge management and suggests the construction of a hierarchy according to importance.
2. **Approach that utilizes Technology Acceptance Model (TAM)** - regards knowledge management system as innovation and examines the behavioural intentions of users to accept this innovation.
3. **Approach utilizing the concept of Gamification** - leverages from the structure of game elements and explains their effect on knowledge management adoption.

2.1 Critical Success Factor approach

The goal of this framework is to determine drivers that systematically predict the knowledge-base acceptance among single users or firms. Such extraction of important factors that impacts the effective functionality and adoption of

knowledge management have been studied in the number of areas and from different perspectives, such as inter-company or intra-company level. The Critical Success Factor concept employed in small and medium-sized enterprises was employed in Wong & Aspinwall (2005). Authors used data from postal surveys to analyze the hierarchy of eleven factors affecting the adoption. These factors were extracted using review of studies rooted in what "early adopters", i.e. large companies, were doing to take advantage of their knowledge. In the next step, the respondents of postal questionnaire were asked to rank the factors according to importance. The unit of analysis used here was the organization, thus, single form approach rather than multi-form one (postal survey was answered once by company as a whole and not by every manager in a firm) was followed. The first place in the final ranking of critical success factors was encroached by management leadership and support. The management thus, should promote co-operation and knowledge sharing across company and also provide support to initiate and sustain effort of employees to create content. The second place belonged to culture of the company. This means that knowledge-oriented cultural foundation determined by trust, collaboration and openness is more important than deployment of KMS. Moreover, result suggests that management and firm's culture, that create company's environment and that determines the willingness of employees to participate in knowledge accumulation (Leeson 2013), is an important critical success factor.

Suresh (2013) investigated factors affecting adoption of knowledge management system in various Indian industries. The methodology used resembles the previous study of Wong & Aspinwall (2005) and differs in subject matter, which is in this case middle and top level managers in a firm instead of a single organization (multi-form approach). The results are ranked according to the quality of success and in detail describe all the elements engaged in a knowledge management system acceptance process. Recognition of knowledge and organization culture were placed in the top of the hierarchy and are considered to be certainly more important predictors of adoption than deployment of KMS technology. Suresh (2013) identified components of these factors for better understanding of how they drive knowledge acceptance within a company and such components were submitted into the questionnaire. The above mentioned recognition of knowledge thus, includes for example recognition of employee's contribution towards knowledge management (firm should attract and retain talented people who are able to deliver good knowledge) or knowledge sharing that firm induce by making contents of knowledge base available.

The other factor, organizational culture, is determined by knowledge-intensive environment, collaboration, emphasize on knowledge sharing and trust. The final ranking of factors divided into components provides a deeper analysis of drivers for knowledge management adoption, practice and innovation.

Yeoh & Koronios (2010) employed critical success factor approach to study business intelligence systems successful implementation. He argues that critical success factors applicable to other types of information systems may not necessarily apply to a contemporary business intelligence system.[1] In contrary to previous studies he thus utilized different method for critical success factors and success measures extraction was applied on five different organizations (cases). According to his findings, system use is (in addition to system quality and information quality) one of the three measures that determine successful implementation of business intelligence systems. Moreover, he indicates that organizational and process-related factors are more influential than technical factors.

To analyze drivers that influence knowledge accumulation in knowledge base it is vital to look at the behavior of knowledge creators, users that systematically interact within installed system. So far, authors employed companies (or executive officers per each company) as unit of analysis (Yeoh & Koronios 2010; Suresh 2013; Wong & Aspinwall 2005). On the one hand, this approach provides an insight from unit that controls all the processes and thus understands the application of knowledge management system. On the other hand, knowledge base is a collaborative product conducted by workers that are willing to share their precious knowledge hence, analysis within a firm instead of analysis between firms is needed. Abril (2007) applied this approach and studied adoption of KMS through behavioral model aimed on workers in a single corporation. Using the shadowing and action research he identified following drivers of behavioral change towards KM adoption: personalized value, executive sponsorship, enabling support organization or incremental perceived success. By personalized value the author means that if managers who are responsible for hiring employees are perceived about value of knowledge-base adoption then employees' cultural change towards the adoption would be induced. Executive

[1]Business intelligence can be defined as "a collection of tools and methodologies that transform the raw data that companies collect from their various operations into useable and actionable information" (Kaula 2015). According to Yeoh & Koronios (2010), implementing a business intelligence system is not an activity that includes the purchase of software and hardware but it is a complex adoption requiring appropriate infrastructure and resources over a lengthy period.

sponsorship also induce adoption but this time by means of inclusion of knowledge management objectives at the leadership team. Enabling support organization driver represents creation of collaborative environment for teams. And finally, incremental perceived success assures that knowledge management system has to be perceived to be successful to affect such behavioral change. This study provides complex outlook to the day-in-a-life storyboards of employees and explain their motivational aspects to participate in KMS. However, users' interaction via knowledge base is not captured, and study lacks this deeper insight into firm's processes.

2.2 Technology Acceptance Model

The Technology Acceptance Model was introduced by Davis (1992) to explain why a user want to use technological innovation. These individuals' intentions are determined by two beliefs: perceived usefulness defined as the extent to which a worker believes that the use of a particular system would increase his job performance and the second, perceived ease of use, defined as the extent to which a user believes that using such a system will be free of effort. In this sense, knowledge management system is considered to be an innovation in a company and these studies examine the factors that lead workers to accept this innovation. Huang & Lai (2014) utilized the technology acceptance model approach to study the effects of three factors on attitude toward knowledge management adoption: perceived usefulness, complexity of the system and the subjective norm defined as perceived pressure or expectations of the community that affect the decision to engage or not to engage in a certain behavior. Author found the positive relationship between perceived usefulness and technology acceptance and also between the subjective norms and behavioral intentions to accept the technology. In case of complexity, the relationship with users' attitude to accept knowledge management was proved to be negative. Ritchie *et al.* (2011) employed the technology acceptance model and analyzed influences of perceived usefulness and perceived ease of use on the behavioral intentions. He states that user acceptance of knowledge management system depends not only on a technology acceptance but also on the organizational and cultural influences.Technology acceptance model was also utilized by Kuo & Lee (2011) who studied effects of perceived usefulness and perceived ease of use on users' behavior. Additionally, he determined also compatibility factor to be important in a sense that if use of knowledge management system is compatible with

the work practices of the users, it also enhances their intention to use the KMS. Using structural model and principal component analysis he confirmed the positive relationships between behavioral intentions to accept knowledge management system and all three factors.

Hou (2014) investigated determinants of user acceptance of business intelligence systems using technology acceptance model. Additionally, he employed its extensions that considers also attitudes, subjective norms and perceived behavioral control. According to his findings, the important influences on users' behavioral intention to use business intelligence systems are employees' attitude cultivation and subjective norms. Thus, both peer opinions and managers' appreciation of successful use of business intelligence platform may motivate users to use this platform.

Regarding the same implications as in technology acceptance model, that process of knowledge management system implementation can be considered as a process of innovation, the interesting results can be found in paper written by Gopalakrishnan & Bierly (2001). In this study author examines impact of three innovation types based on dimensions of knowledge on innovation adoption.[2] The results suggest that the more tacit (unable to codify or articulate) and complex knowledge associated with innovation, the higher level of innovation adoption is reached.

The goal of studies utilizing technology acceptance model as well as critical success factor approach is to determine factors that should be emphasized in order to enhance adoption of knowledge management system. The findings of such studies serve as a systematic guidance for companies according which they might direct their management. For example, Suresh (2013) highlighted sharing knowledge as one of the important components of critical success factors. Following technology acceptance approach, Hou (2014) identified employees' attitude cultivation as a driver of behavioral intention to use knowledge management system. Although, these results define functionality of knowledge-base platforms they lack deeper explanation of how such factors can be used to motivate users to create content. In other words, knowledge-base creation is in hands of knowers (Davenport *et al.* 1989) thus, analysis of drivers that affect users' motivation and behavior towards collaboration should be emphasized rather than general firm-level factors. The following section thus offers the new

[2]The dimensions of knowledge are tacit-explicit, systematic-autonomous, and simple-complex (Gopalakrishnan & Bierly 2001).

concept, called Gamification, that might be capable of influencing productive behaviors of users (Leeson 2013).

2.3 Gamification

In an innovative paper Leeson (2013) argues that the culture is a lynchpin that will determine the workers collaboration and system adoption and that the valuable tool to encourage this process is a new concept, called gamification.

2.3.1 Gamification Concept

Hamari (2013) defines gamification as a process in which services are enhanced with motivational stimulus in order to invoke gameful experience and further behavioral outcomes. Deterding & Dixon (2011) provides simpler approach and define gamification as "the use of game design elements in non-game contexts". It is a new term for relatively old method. One example might be education and its gamified approaches from Scrabble used to teach spelling to duoLingo - application for learning languages.[3] We can also find its main characteristics in strategies that are used to maintain customers interaction like Customers Relationship Management including loyalty systems, etc. (Ballance 2013). In practice, only one part of games is incorporated in non-game contexts - scoring.[4] Users of gamified systems are thus motivated to use such system more by obtaining points, badges or reaching leader-boards and higher levels. Since new technology era, principally era of smart-phones and tablets, gamification is strongly connected to social interaction (likes or dislikes from other users/players, etc.). Hence, gamified experience brings not only feeling of self pride (by reaching leader-board or more points in some activity) but also satisfies the need for socializing (Moise 2013). Gamification then seems like reasonable approach for motivating knowers/employees to deliver and seek further content.

2.3.2 Gamification in Knowledge-Base Adoption

In theory, gamification can be divided into three parts: 1) the implemented motivational stimulus, 2) the resulting psychological outcomes, and 3) the further

[3] `www.duolingo.com`

[4] Nicholson (2012) suggests "pointsification" as a label for gamification systems that add nothing more than a scoring system to a non-game activity.

behavioral outcomes (Hamari & Sarsa 2014). Leeson (2013) suggests that the correct combination of game mechanics and behavioral economics may lead to long run increase in users' intention to accept KM and share their knowledge. However, he emphasizes that the most important issue is to direct employees to realize the inherent benefits of collaboration via boosting their intrinsic motivation instead of the extrinsic one.[5] Thus, even the introduction of such gamified tools as badges or leaderboards in knowledge-base platform can lead only to short run change in users' behavior. This statement is also supported by Nicholson (2012). In his paper, he claims that rewards can reduce internal motivations as firm which temporarily implements external payoff system will be after quitting such a program worse off than before implementation. Users will be simply less likely to return to the behavior without the external reward.

The working idea utilized in the paper by Leeson (2013) is approach proposed by Pink (2009). He argues that human motivation is largely intrinsic and he identifies three powerful ways to induce this kind of motivation: autonomy, mastery and purpose. Autonomy allows users to set their own goals and to control their activity. The more free is a knower to decide how to collaborate (write a comment, thank for created page, etc.) the more he will be engaged in sharing a knowledge. Mastery is about obtaining a good skill in something which yields own inherent benefits. And finally, purpose ensures the social connection to the larger entity via the channel of making a broader impact. Collaboration within a knowledge management system therefore, leads to a higher purpose - further creation of a collective wealth of information and experience.

Although, the concept of gamification is new and is still evolving, some studies has analyzed effects of implementing the gamified tools into knowledge management system on further content creation. Farzan (2008, 2008a) in his work utilized the system of points in networking website for employees and studied the impact on their collaboration. In his framework, he divided users into two groups. The experimental group which was rewarded by points if engaged in the knowledge creation and the control group that was not rewarded and did not know about point system. The framework has several elements, but the most important is the idea that the user from control group (without rewards) is in the long run motivated by the higher activity of other group

[5]Intrinsic motivation happens when people engage in activities for the activity itself and without any obvious external incentives such as rewards. Extrinsic motivation happens when people engage in activities as a result of an external incentive mechanism such as contingent rewards (Farzan & Brusilovsky 2011).

members and also by previous activity of experimental group. In particular, results of both studies showed that the point system does increase the knowledge creation of the experimental group and that their higher activity serves as an intrinsic trigger for control group participation in the next period. Moreover, Farzan *et al.* (2008) states that gamified tools installed in the knowledge base stimulate the discussion among users and that workers' action depends on what others do. Farzan & DiMicco (2008) in her study utilized data in form of a log into the database, where every action of all users is recorded, so as an independent observer can analyze which activity within a studied system contributes to content creation. The main idea is that, this method of data accumulation provides detailed insight into creation of each component of knowledge base and hence, allows studying the incentives' characteristics.

In his next study, Farzan & Brusilovsky (2011) employed new incentive scheme installed in community-based course recommended system that provides personalized access to information about courses and which turns user participation into a self-beneficial activity. Users are here provided with incentive scheme that motivates them to collaborate and rank courses. In particular, the users evaluate the relevance of each taken course to each of their self-selected career goals. When subsequent users are choosing courses they can decide according to the degree of relevance toward their goals and hence, contributors' activity is beneficial to the community as a whole when users engage for the activity itself. Students are then supposed to be motivated by the tool that shows their progress towards their self-selected goal. This was followed by subsequent analysis in which the positive relationship between working mechanism and users' collaboration was found. Nevertheless, author emphasizes the problem of self-deception that can cause the higher rating by students who want to attain a higher visible progress. The study thus, hints the deep consideration of the incentive mechanisms used, as effect of extrinsic motivation on intrinsic one can raise the possible drawbacks. Author further argues that in both large and small communities, the most important issue is to motivate the largest percentage of users possible to contribute. While small knowledge-management-systems' survival depend on contribution of majority of users, larger communities (like Wikipedia) with large amount of users is able to survive with small percentage of contributors. However, even such big knowledge base can suffer from participation inequality bias problem when

small percentage of users represents the views of larger population.[6]

Finally, the connection between possible implications of gamified knowledge base and behavioral economics was offered by Hamari (2011). He suggests that concepts utilized in behavioral economics can be used to explain the effects of the certain game design patterns installed directly in knowledge management system. The main concept used in this study is loss aversion in connection with prospect theory, according which losses loom larger than corresponding gains (Kahneman & Tversky 1979). This framework of decision making under risk that systematically violate the predictions of expected utility theory has been found in decision making in different areas like consumption-savings decisions, labor supply or insurance (Barberis 2013). Hamari (2011) further suggests sunk-cost fallacy theory (Arkes & Blumer 1985) to explain potential intentions why users participate in further content generation. Along the lines of this theory, people are far more willing to invest to the activity that they have already invested in. Therefore, supposing risk aversion of users and an assumption that a proper incentive scheme is in place, users participate because they have already participated before. Adoption of knowledge base thus depends on users' previous activity.

[6]As of 2008, Wikipedia had 684 milion unique users, while only 75 000 (0.01%) of them actively contributed (Farzan & Brusilovsky 2011).

Chapter 3

Design and Elements of Knowledge Bases

Before we introduce the framework of our study it is essential to describe how knowledge bases designed by Semanta work and what are their elements. As discussed in Chapter 1, knowledge management systems employed in our analysis are available to workers through common web browser. They are applications based on user-generated content similar to Wikipedia, consisting of a huge number of pages organized into trees and hierarchically ordered in classes and subclasses. Users of these systems are able to see how pages are related and can be navigated to other pages using links. The most important characteristic of such systems is that within them pages can be easily created or edited. This allows users to collaborate and continuously create content compared to the usual web pages that can be only visited without possibility to contribute to them.

Knowledge management systems are not typical open-sources as they are open only to individuals with granted permission - usually employees or other external workers. However, they incorporate many elements used by popular websites (Wikipedia, Facebook, LinkedIn, etc.) used to share individually produced content. There are dozens of such elements and features which change and evolve over time. Therefore, we will concentrate only on those which are most important and most widely used. These elements are: creating pages, editing pages, commenting pages, *Thank you* buton and *Ask* button.

- *Creating Pages* is within analyzed knowledge bases performed by button *Add Page* that can be found in all system pages from Home Page to the last page in tree-like hierarchy. This means that users are able to

create their own content by placing their pages anywhere in a tree directly specifying its relation to other pages (parent page, child page, etc.). The process of creation is done through automatic form that is displayed after user clicks on *Add Page* button. The form requires specification of page title and insertion of the content (text, table, picture, figure, attachment, etc.).

- *Editing Pages* is accessible using *Edit* button. As in previous case, this button is part of every page in a system, however, some of them might be restricted and can be edited only by some employees (for example, pages containing important information on suppliers can be edited only by administrator). In such situations *Edit* button is still present but is not active and after clicking on it the edit form is not displayed. Thus, if an user is allowed to edit some page, and she clicks on the button, automatic form appears and is pre-filled by the page content. User can rewrite it, add new passages, insert or delete tables, graphs, attachments, etc. Any such change in original content of the page is considered as edit (even if an user only corrects the grammar).

- *Commenting Pages* is an element mostly known from social networks. As well as any content inserted in these networks (blogpost, photos, etc.) also pages in knowledge bases can be commented by other users. Users can find pre-inserted box at the bottom of every page, fill it and click on *Send* button to save it. Comments are then immediately showed on the given page. Comments are usually created by other user than the one that generated the page while edits are usually performed by page-creator herself.

- *Thank you* button is again element well known from social networks. Using example of Facebook, this button is similar to *Like*. The button is placed in the bottom of every page and after user clicks on it, the button changes the color and the information that the page was thanked appears next to it.

- *Ask* button is present in the bottom of every page next to *Thank you* button. By clicking on it, user is provided with a form in which she can specify a question or request and assign it a title. By submitting this form the question is directly sent to a relevant expert and is saved as a new page in a special section of knowledge base designed for requests. This question

is then answered by inserting comments to its corresponding page so as it is visible for everyone. Alternatively, the question is answered using "The Best Answer" box by editing the question-page.[1] Moreover, not only experts but also other co-workers are allowed to join the answering process and insert comments to such pages. This element is very important as it allows users to ask for knowledge when they can not find it or when it is simply missing.

[1]To make it clear, suppose we would like to ask the following question: "What is this thesis about?". We will fill the request form (predetermined by the system) which will allow as to state our question and to specify for example title of a question (let's make it "Thesis") or the field of a question (let's suppose "academic"). This will directly create a single page (our question) in the space dedicated to requests and will automatically notify the expert in "academic" field about new item added. The expert (or any other system's user including the asking person) is able to comment this page or to edit this page (only the expert or the admin) by filling the special prearranged box "The Best Answer". The both actions produce the response to the question.

Chapter 4

Hypotheses

The thesis estimates the effect of various activities within a knowledge base, performed by users of respective system on organizational knowledge-base adoption. Following Kuo & Lee (2011), knowledge management system is adopted if users continuously share and seek knowledge within it. By sharing knowledge via knowledge management system, users convert their own personal knowledge into corporate one - they are creating knowledge base. By seeking knowledge they are extracting benefits of corporate knowledge which leads to facilitation of users' task execution (Suresh 2013). Thus, the two components of adoption can be stated as:

(i) continuous creation of knowledge-base's content, and

(ii) continuous knowledge-seeking.

4.1 Hypothesis #1

According to Farzan & DiMicco (2008), user's content production within knowledge management systems is enhanced by activity of other users. As discussed in previous chapters, we assume three users' actions leading to content generation: creation of pages, editing of pages and inserting comments. The nature of KMSs analyzed allows users to visit the created content and collaborate on it by commenting or thanking the creator. We assume that such activity of co-workers interacting with the content creator positively affects creator's intentions to generate more content. In simple words, if co-workers are visiting creators pages, or if they are commenting it or giving "thanks", the creator should be motivated to add more pages into knowledge base (as she assume

her content to be important to others). Alternatively, depending on nature of comments, creator might be motivated to edit the page in order to correct or elaborate more on ideas, etc. Moreover, broader knowledge base (as for amount of content) provides more opportunities for collaboration and thus, we also expect that knowledge-base size positively affects users' contributions. So, we formulate the first hypothesis as:

Hypothesis #1: *Further content creation (creation of pages, edits of pages and comments) depends on collaborative activity of other users - page visits, page comments, thanks for pages as well as on knowledge-base size.*

4.2 Hypothesis #2

Users' collaboration and knowledge sharing is an essential determinant of successful knowledge-base adoption. To study the effects of users' activity on sharing knowledge (and thus, on knowledge-base adoption) we will also employ the gamification concept (Section 2.3). This construct uses game elements (those that make games engaging and attractive for "players") and apply these components in other contexts (Leeson 2013). It offers an answer on how to promote desirable users' activity within a system. In other words, gamification is a solution, in which content of knowledge base is created because users are motivated to contribute and collaborate by "gamified" tools directly installed in knowledge management system. This thesis considers such tool to be leader-board-like Hall of Fame placement.

4.2.1 Hall of Fame Page

The Hall of Fame represents a single page in a system that serves as an information portal about top 5 positions according to categories in workers' collaboration. Activity of all users is here evaluated and any viewer of Hall of Fame page can see chart of people who dominated in given category in a previous week. Tracked categories are:

1. *contributor* - sequence of maximum of 5 users who created and edited the highest amount of pages in the previous week,

2. *commenter* - sequence of maximum of 5 users who commented the highest amount of pages in the previous week,

3. *consumer* - sequence of maximum of 5 users who visited the highest amount of pages in the previous week,

4. *Thanks receiver* and *Thanks giver* - sequence of maximum of 5 users who obtained or gave the highest number of thanks to any content created (see Chapter 3).

The positioning is recalculated in the beginning of each week. Hence, in every Monday there are new scores regarding activity in previous week.

In such setting we expect that knowledge base is adopted (more content is shared) as a result of users' interest in chart-leading placements (visiting Hall of Fame page) and employees' natural behavioral intentions induced by gamification. We suggest it happening in two directions. Firstly, employees that know that they reached some position in the Hall of Fame page want to defend and keep it also for next period. They will create more content so as they appear in the page also in one week. Secondly, if users find out that their activity was not sufficient to be part of a Hall of Fame, they would try to beat others so as to appear in the leader-board next week. As an implication, we expect that not visiting Hall of Fame page affects users' activity in lower or no extent. We can formulate the second hypothesis as:

Hypothesis #2: *Further content creation (creation of pages, edits of pages and comments) is promoted by gamified tools, concretely by viewing placements in Hall of Fame page and by previous positions reached in the Hall of Fame leader-board.*

4.3 Hypothesis #3

Another part of our analysis is dealing with estimation of effects on knowledge seeking as an important factor of adoption. If users continuously search for new contextual information they adopt the processes incorporated in knowledge base. It might happen that in a given point of time a certain knowledge is missing. Therefore, users are able to ask questions and request an expert's insight in order to obtain such valuable peace of missing knowledge. We expect that user will be motivated to seek knowledge more frequently (will visit the system more in one-week period after she obtains answer to question) if the expert provides fast response to such request. We also assume that knowledge seeking is boosted if user is contented with the answer. An employer's satisfaction is the outcome of how her request was treaded and to what extent an

expert was involved in the response. In other words, the number of answers and variety of these answers (number of different experts dealing with it) affects knowledge seeking. We can thus, formulate the third hypothesis as:

Hypothesis #3: *Knowledge base adoption (knowledge seeking) is driven by speed of response to questions and requests and by variety and amount of these answers provided by system experts.*

In the next sections we will describe the data that will serve for these relationships' estimation and we will introduce the methodology.

Chapter 5

Data

5.1 Knowledge Management Systems

The analysis throughout this thesis is conducted by utilizing the datasets from two respective systems administrated by Semanta, s.r.o.[1], Semanta Guides system and system designed for one of the mobile operators operating in the Czech Republic. Guides is knowledge management system created for collaboration with Semanta's partners. Its users are thus the partners (employees of partner firms) and Semanta's internal employees. It represents a knowledge base that contain documentation related to all Semanta's products and services, methodologies, how-to procedures or training materials. It is a place where users are able to find any information regarding Semanta and its processes, collaborate on them, collect feedback, comments and suggestions or request information. The second knowledge base is used by employees of mobile operator or its external co-workers. It is a big corporation and this nature affects also number of registered users that is at least ten times larger than in Guides system. This knowledge base contains internal information about projects, marketing, sales and other areas of operator's interest. Employees are able to collaborate, share knowledge about their experience, ideas and insights and also ask experts for missing knowledge. The reason why we decided to include two systems for testing our hypotheses is to control for possible selection bias (each knowledge base differs depending on community of workers creating it). Throughout this text we will use the following abbreviations:

- **Guides** for Semanta Guides system, and

[1] `http://semanta.cz/home.html`

- **MO** for system utilized by above stated mobile operator.

The next parts will show the same structure of both Guides and MO data. However, the studied knowledge management systems operate in different environments. While Guides system is employed in the small firm, the MO runs in the big corporation with different (more regulated) personal organization and more formal operation. Table 5.1 shows that the total number of registered users in a Guides system is more then twenty times lower than MO and average number of weekly visits to MO system is almost four times higher than to Guides. However the average weekly activity (number of pages, comments and edits created by all users) exhibit opposite tendencies. This suggests that Guides knowledge base is more frequently used for content creation although there is considerably greater amount of registered users and resulting system's visits in MO. Thus, we will perform separate inference for Guides and MO data although, our hypotheses hold for adoption of a broad range of knowledge bases.

Table 5.1: Basic Statistics for Guides and MO

	Guides	**MO**
No. of Registered Users[a]	225	5666
Average Pages[b]	23.75	11.36
Average Comments[c]	43.05	0.51
Average Edits[d]	129.67	45.29
Average Visits[e]	994.47	3908.364

Notes:

[a] Total number of ever registered users in respective systems as in June, 2015

[b] Average number of pages created by all uses in respective systems during the week. Data on weekly activity comes from year 2014.

[c] Average number of comments created by all uses in respective systems during the week. Data on weekly activity comes from year 2014.

[d] Average number of page-edits performed by all uses in respective systems during the week. Data on weekly activity comes from year 2014.

[e] Average number of users' visits of respective systems during the week. Data on weekly activity comes from year 2014.

5.2 Capturing the Data

Semanta collects data on every action that takes place in its systems. The action is considered to be any click performed by a user within any system's

page that leads to the realization of some event. These events might be for example: creating pages, editing pages, inserting comments, asking using *Ask* button, thanking for a content using *Thank you* button, etc. (see Chapter 3) or events connected to viewing pages (as user has to click on some link navigating him to the page) and also log in and log out events.[2] Each such captured action is stored as a row in a table indexed by respective id. This row then contains information on *who, when* and *where* performed *what*. Hence, some row in a table might display information that user A (*who*) inserted comment (*what*) in a page created by user B (*where*) on June 6th, 2014 (*when*), etc. The tables thus, represent history of everything that was done within a knowledge bases allowing us to extract data capturing users' activity and collaboration and also data on certain actions performed within a system.

In the previous chapter (Chapter 4) we specified two important components of knowledge-base adoption by users that we are studying in this thesis: continuous content creation that will be analyzed in first two hypothesis and continuous knowledge seeking which we analyze in third hypothesis.

5.2.1 Data Extraction - First and Second Hypothesis

We have defined content creation as creation of pages, editing pages and commenting pages. The aim of this study is to estimate effects of selected factors on the amount of these actions performed by a user in some period of time. In both hypotheses we set this period to be one week as we consider this period to be sufficiently long for monitoring users activity. To extract variables capturing amount of the actions of our interest (performed by each user in each consecutive week) we have firstly selected rows corresponding to the given event (*what*). Then we grouped this selection according to users (*who*). Because every row contains information on time when the action was performed (*when*), we were able to determine which rows correspond to which week. Our final step was then counting rows matching every possible user-week combinations. Hence, every variable resulting from this process consists of three columns: *user_name* identifying a studied user (*who*), *week_code* determining exact week (*when*) and frequency of given action (creation of pages, edits or comments) performed for each user-week pair. These variables are used as dependent measures in our first and second hypothesis.

[2]There is a wide range of other events/actions that are targeted in Semanta's data-collection however, these actions are not concerned in our analysis and hence we will not discuss them.

In Section 4.1 we introduced factors affecting content creation that will be studied in our first hypothesis. Those are collaborative activities of co-workers - their visits to user's pages or their comments and thanks to user's pages. From the above stated table, we were able to extract variables representing how much of these co-workers' activities happened in interaction of each user (concretely, her pages) in each consecutive week. In other words, if we take visits-to-user-page factor, the extracted variable represents number of co-workers' visits to pages created by each user in each consecutive week. The process of extraction partially resembles the previous one. We started with selecting rows corresponding to the given event (*what*) - visiting pages, commenting pages or thanking for pages. Then we grouped this selection by a user who created the page which was visited, commented or thanked for (*where*). Again, using the *when* information we determined which rows correspond to which week and finally, we counted the rows matching the user-week combination. Factor variables are then of the same structure as dependent measures described above and contain three columns: *user_name* identifying creator (*where*), *week_code* (*when*) and frequency of given co-workers' activity. Although these factors were not stated in Section 4.2 introducing our second hypothesis, we will employ them in regression as control variables. Other explanatory measures used in the first and second hypothesis follow from the similar extraction processes. Therefore, we will not provide their description in detail.

While every extracted variable contains two columns (*user_name, week_code*) that are identical across these variables, we were able to merge the measures based on those columns. As a result we obtained panel data with *user_name* specifying panel variable and *week_code* standing for time variable. Although dependent measures employed in first and second hypothesis are extracted using the same process, observation periods for these hypotheses differ and do not coincide. Moreover, Semanta started to collect data needed for our second hypothesis in February, 2015 (five months before completion of this thesis) and this observation period was not sufficiently long for less frequently used MO system to be appropriate for analysis.[3] And finally, data resulting from the analyzed knowledge bases, Guides and MO, are collected in separate initial tables. Therefore, we will employ two datasets in case of first hypothesis and one dataset in case of second one.

[3]After plugging the data into the model, regression analysis was not computationally feasible.

5.2.2 Data Extraction - Third Hypothesis

Knowledge seeking considered to be the second important component of knowledge-base adoption, is defined as a process of visiting system's pages. Instead of analysis of actions performed by each system's user in every week, in our third hypothesis (Section 4.3), we concentrate only on activity (system visits) performed by users who asked the experts for missing knowledge (using *Ask* button, Chapter 3). The aim of the hypothesis is then to estimate effects on a number of visits performed by these users in the week after they obtained answers. The extraction of variable capturing this was again made using tables collected by Semanta. Firstly, we have selected the rows corresponding to action - asking for knowledge. We obtained rows representing a list of every asked question within the system (1). These rows contained information on which user (*who*) asked the question and when it happened. Because every asked question is transformed into page after submitting it, and answering a question means inserting a comment into it (Chapter 3), we found the dates of answer (*when*) as follows.[4] In a first step, we selected rows corresponding to comment-page action with the condition that *id*s of pages in which the comments appeared (*where*) have to match with the corresponding *id*s of pages from (1). We obtained rows representing list of all answers to questions from (1) and thus, also information on answer dates (*when*). However, some questions might have been answered more than once. Hence, in the second step, in case of multiple-answer to a question, we chose only the row with the minimum date. In this point, we know *who* asked the question and *when* he obtained the answer. Setting the period of our interest to be one week after an user (*who*) obtained answer we can extract the frequencies of these users' visits (*what*) by using the same procedure as in Subsection 5.2.1. Importantly, the resulting dependent variable is not composed of three columns as in previous two hypotheses. It contains only above specified event counts because we are not concerned with exact user or week connected to it. The extraction of other variables employed in third hypothesis is done using similar procedure as we described for dependent measure. Resulting datasets for both, Guides and MO system, is thus, structured into cross-sections of questions that were once asked by an user whose activity (further knowledge seeking) is then subject of our

[4]The questions might be also answered using "The Best Answer" box as described in Chapter 3. The process of obtaining data on this type of answering is the same as in case of answering with comments. If a question is answered using both mentioned types, then answer date is selected as the minimum among all answer dates arising from the process.

analysis.

5.3 Data for the First Hypothesis

Hypothesis #1: *Further content creation (creation of pages, edits of pages and comments) depends on collaborative activity of other users - page visits, page comments, thanks for pages as well as on knowledge-base size.*

In previous section we described extraction of the data for our first hypothesis that resulted in two panel datasets, one for Guides system and second for MO system, in which users are observed over one-week periods. Data are of balanced-panel structure since we can observe each user in every week in both systems. MO knowledge base accounts for longer history of monitoring than Guides hence, we decided to use different periods of observations for the two analyzed systems:

- in case of Guides, from January 2014 till February 2015, that results in 62 one-week periods for each observed participant, and
- in case of MO, from June 2012 till February 2015, that results in 121 one-week periods for each observed participant.

The above stated balanced panels (all observations for each user in every period are included in a sample) predicts zero events' observations when a given user does not perform any activity within a system during one-week period. In other words, if a user does not create any page, edit or comment in a given week, the resulting values of variables capturing it are zero. The problem arises when for a certain user there is a considerable amount of zero observations and thus, the sum of frequencies of her activity during the whole observed period is very small. The main reason for such a low participation is only a temporary access to system given to some users. Once such participant perform any type of event during period of our study, her activity is immediately captured in our data. Thus, we decided to exclude all users for whom the sum of frequencies of her activity (creation of pages, commenting or editing) during the whole studied period is less than 30 (Guides) or 60 (MO) - only individuals with at least one activity in two weeks on average are included.[5] As a result, we obtained panel structure of data consisting of 15 users in case of Guides data

[5]The two-week period was chosen as a sufficiently long time for absence in content creation in case of vacation, illness or other type of non-presence in system use by workers.

and 11 in case of MO. The final number of user-week combinations is 930 and 1.331, respectively. Summary for this two datasets is shown in Table 5.2.

Table 5.2: Datasets' Summary for First Hypothesis

	Guides	**MO**
No. of one-week periods	62	121
No. of studied users	15	11
No. of obs.	930	1331

However, the examination of dependent measures in the next section suggests overdispersion and excess zero character of data. Taking into account these problems, we decided to utilize complex econometric model: zero-inflated negative binomial. Because its methods are not implemented in common statistical packages (*R* and *Stata*), we decided to follow Allison & Waterman (2002), and will treat these data as cross-sectional by adding dummy variables to estimate individual fixed effects. While we are dealing with low number of studied users and high number of time periods we can apply it on our data. The process is described more in detail in Chapter 6. In the following sections, we will first study properties of dependent measures and show their undesirable characteristics, then we will proceed by describing the explanatory variables.

5.3.1 Dependent Measure: Further Content Creation - First Hypothesis

To study the effect of users' activity on knowledge sharing we will employ three different measures of a user's activity in a given week as dependent variable:

- *create_page_count* - that represents number of pages created by user during one-week period
- *comment_page_count* - that represents number of the user's comments during one-week period, and
- *edit_page_count* - that represents number of page edits by given user during one-week period.

Further content creation is thus, result of each of the three measures.

Our dependent variables represent event counts ranging from zero (if user was inactive in given week) to some positive count. Therefore, we will perform

three distinct regressions, each employing one of our dependent variables while using the same explanatory measures. It is of our concern to investigate to what extent each activity is promoted by contribution of other users and the overall effect would be assessed only qualitatively. Summary statistics for all three measures for both Guides and MO data are presented in Table 5.3.

Table 5.3: Summary Statistics for Dependent Measures - First Hypothesis

	Obs.	**Mean**	**Var**	**Min**	**Max**	**CF20 (in%)[a1]**	**Zeros (in%)[b1]**
Guides							
create_page_count	930	1.4	22.73	0	92	98.92	62.69
comment_page_count	930	2.41	45.18	0	63	97.53	66.13
edit_page_count	930	8.09	221.66	0	135	88.17	37.85
MO							
create_page_count	1331	0.99	9.39	0	36	99.55	75.51
comment_page_count	1331	0.45	3.78	0	33	99.85	85.27
edit_page_count	1331	7.18	380.49	0	190	89.19	65.06

Notes:
[a1] Cumulative frequency up to 20 counts.
[b1] Proportion of zeros in a given dependent variable.

In our data, we are dealing with two main problems connected to event counts (Cameron 1999):

(i) *overdispersion*, and

(ii) *excess zeros.*

(i) Typically, event count data are analyzed using Poisson model. The underlying Poisson distribution assumes that the expected value of the dependent variable is equal to its variance. This is called equidispersion. However, in our data we can see that the mean of all three variables, in both Guides and MO data, is several times higher than its respective variance (in case of Guides and *edit_page_count*, more than 27 times) thus, we can not assume equidispersion.[6] We can also observe several outlying observations as the distribution of all three dependent measures shows long right tail. Cumulative frequency up to 20 counts reaches 98.92% (Guides) and 99.55% (MO) in case of *create_page_count*

[6]The problem of overdispersion might arise due to number of reasons, such as unobserved heterogeneity, outliers, or because the process generating first and the later events may differ Greene (2012).

(97.53% and 99.85% in case of *commment_page_count* and 88.17% and 89.18% in case of the third measure). The later case means that only 1.08% exceeds 20 page creations within a week while the maximum number of pages created is 92. There are three general ways how to treat outliers (Ghosh & Vogt 2012): (1)keep them in sample, (2)winsorize it (assign it lower weight) or (3)drop it from sample.

While we assume the dependent variables to be a reaction to previous activity, we will not consider third option (eliminating these points would mean losing the information on users' reaction). Moreover, treating the outliers as any other points in data may cause the estimates to considerably differ from the true population value (Ghosh & Vogt 2012). Thus, we will not utilize the first way. As a result, although it also represents possible danger for proper estimation of parameters, we employ winsorizing. This method includes replacing any data values above chosen percentile of the sample by a value of a given percentile. Thus, outlying observations are not thrown out but are adjusted so as to be closer to other sample points. Because we assume that very high counts of events might be correct observations but may result from a specific situation in a given week (e.g. importing a dictionary into knowledge base means creating a single page for each term and thus significant observations for *create_page_count* variable) and that these situations are not frequent, we place 99th percentile for each dependent measure for winsorization. The counts (values) corresponding to this percentile are stated in Table 5.4. The resulting histograms of *create_page_count* variables are shown in figures 5.1 and 5.2 (distribution plots for *comment_page_count* and *event_page_count* can be found in Appendix A).

Table 5.4: Values (Event Counts) of Dependent Measures corresponding to 99th percentile

	Guides	**MO**
create_page_count	21	13
comment_page_count	37	9
edit_page_count	78	100

(*ii*) Secondly, the proportion of zeros in data is higher than required by Poisson distribution (see Chapter 6). We detected 66.13% of nulls regarding *comment_page_count* variable (Guides) meaning that for 66.13% of observations an user was inactive in terms of commenting pages in a given week. The

Figure 5.1: Histogram of *create_page_count*, Guides

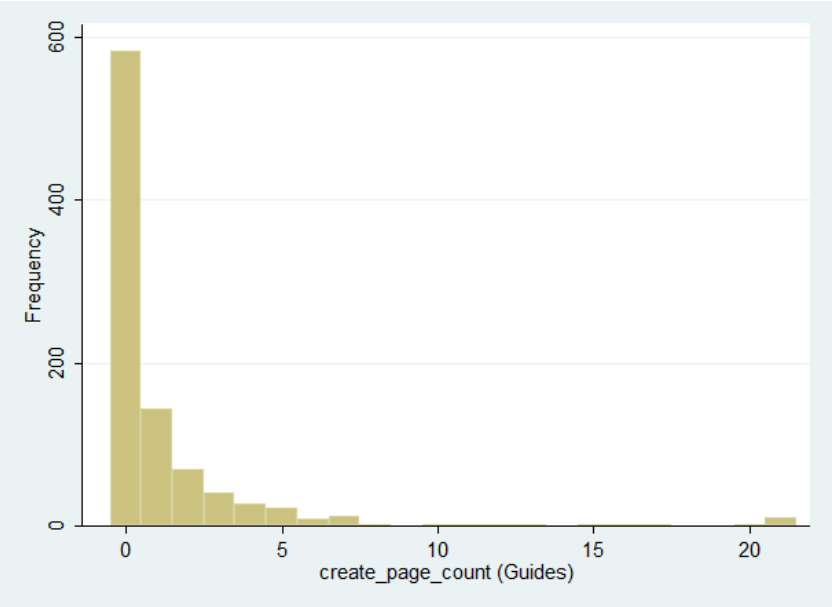
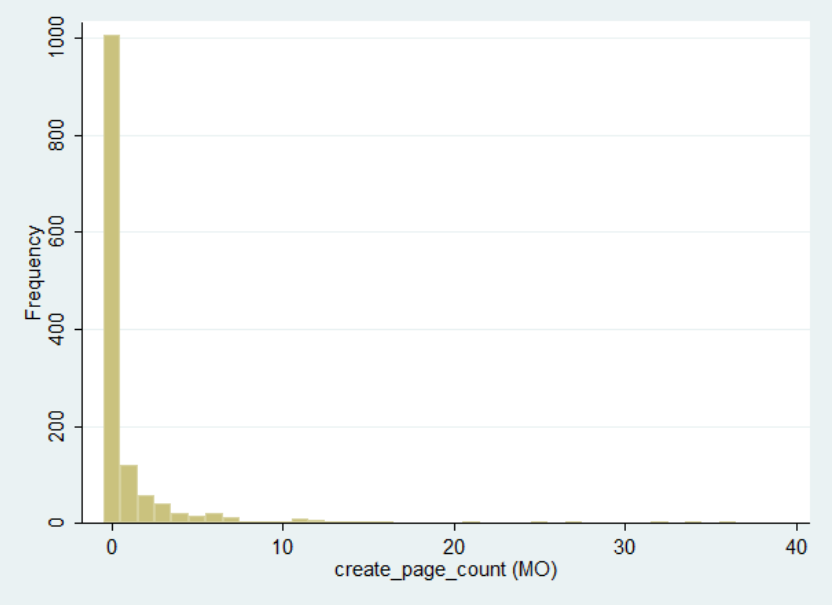


Figure 5.2: Histogram of *create_page_count*, MO



later holds for the rest two dependent measures in Guides system for which proportion of zeros amounts to 62.69% and 37.85%, respectively. The mean of all variables is thus, low because the data indicates high proportion of zeros and substantial cumulative frequency up to 20 activities of given type.

Table 5.5: Detailed Summary Statistics for Dependent Measures - First Hypothesis

	Guides			**MO**		
	create_page_count	*comment_page_count*	*edit_page_count*	*create_page_count*	*comment_page_count*	*edit_page_count*
Obs.	930	930	930	1.331	1.331	1.331
Median	0	0	2	0	0	0
Mean	1.40	2.41	8.09	0.99	0.45	7.18
St. Dev.	4.77	6.72	14.89	3.07	1.94	19.51
Variance	22.74	45.18	221.66	9.40	3.78	380.49
Skewness	10.73	5.11	3.61	6.25	8.62	4.47
Kurtosis	165.74	35.29	20.51	56.22	102.39	28.22

The detailed summary statistic is shown in Table 5.5. All three measures in both datasets are positively skewed as indicated by positive skewness coefficient as well as by fact that means of all variables are higher than their medians. The kurtosis values suggest leptokurtic distributions which reflect acute peak around the mean and fatter tails (in comparison with normal distribution with kurtosis equal approximately 3). These properties identify Poisson distribution (see Chapter 6).

5.3.2 Independent Variables - First Hypothesis

In analysis of both, Guides and MO data, we will use the following variables: *my_page_visits_ count, my_page_comments_count, KBsize* and *Dummy for FE*. In case of Guides, we will include additional variable *my_page_thanks_count*. This measure will not be employed in analysis of MO system because actions in which users are thanking for creators' pages in this system was not monitored during the MO's studied period. The procedure leading to extraction of underlying variables was presented in Subsection 5.2.1. Their definitions are presented later in a text. Detailed summary statistic of all these measures can

be found in Table 5.6. In case of first three variables, these statistics show very similar nature of properties in comparison to independent measures - they are event counts with overdispersion and excess zeros problem. We are also dealing with several outlying observations. Hence, we will follow the same procedure of winsorization as in Subsection 5.3.1 and we replace the values of observations on *my_page_visits_count* and *my_page_comments_count* variables that are above 99th percentile by the corresponding value of this percentile. Table 5.7 summarizes these values. Data on *my_page_thanks_count* are not winsorized as the maximum value of this variable is count of 8 events (thanking for creator's pages) (Table 5.6).

To support our predictions we use all variables presented in Table 5.6 and we assume them to have positive relationship with content creation (except for Dummy for FE variable, whose effect will not be studied). The description of respective explanatory variables follows:

- ***my_page_visits_counts*** - an event count variable that represents count of all visits during the one-week periods to pages created by a certain user - creator. This event count includes only visits performed by other users and not visits performed by the creator. This is because we want to capture the collaborative activities within systems. We assume that this event count positively affects knowledge generation (creation of pages, editing pages or commenting pages) of creators. Results in Table 5.6 suggest overdispersion (variance is more than 230 times higher than mean in Guides and almost 300 times higher in MO) and excess zeros problem (the proportion of zeros in Guides and MO sample reaches 19.78 % and 27.05 %, and cumulative frequency up to 20 visits is 55.81 % and 71.90 %, respectively). Again, due to the balanced-panel property, zeros in this variable result from situations in which co-workers are not visiting creator's pages during a one-week period. The proportion of such observations in *my_page_visits_count* is 19.78 % (Guides) and 27.05 % (MO).

- ***my_page_comments_count*** - variable that represents count of all comments inserted into pages that were created by a certain user - creator, during each one-week period. Again, we are considering only count of such events performed by other users than creator herself to assess collaborative activities. Hence, this variable is of same structure as previous one. However, the frequency of comments is considerably lower than fre-

Table 5.6: Summary Statistics for Explanatory Measures - First Hypothesis

	my_page_ visits_ count	*my_page_ comment _count*	*my_page_ thanks_ count*	*KBsize*	*Dummy for FE*
	Guides				
Obs.	930	930	930	930	930
Mean	57.66	1.39	0.07	7,743	0.07
Median	14	0	0	7,934	0
St. Dev.	116.70	4.35	0.42	1,316	0.25
Variance	13,617	18.93	0.17	1.73 mil	0.06
Min	0	0	0	5,016	0
Max	869	52	8	10,290	1
CF20 (%)[a]	55.81	98.92	-	-	-
Zeros (%)[b]	19.78	77.31	95.48	-	93.33
	MO				
Obs.	1.331	1.331	-	1.331	1.331
Mean	39.87	0.22	-	16,252	0.09
Median	5	0	-	19,633	0
St. Dev.	108.49	1.19	-	7,985	0.29
Variance	11,771	1.43	-	6.38e07	0.08
Min	0	0	-	2,476	0
Max	1,068	29	-	24,075	1
CF20 (%)[a]	71.90	99.92	-	-	-
Zeros (%)[b]	27.05	90.83	-	0.83	90.91

Notes:

[a1] Cumulative frequency up to 20 counts.

[b1] Proportion of zeros in a given variable.

quency of page visits. This can be seen from Table 5.6 showing that in 77.31 % of observations, the Guides-system's users do not comment creator's pages at all (for MO it is 90.83 %). Moreover, in 98.92 % (for MO - 99.92 %) cases, the sum of all comments connected to given combination of creator and one-week period is lower or equal to 20 while maximum count is 52 (29). As in previous case, we expect positive relationship between dependent variables and number of comments to creator's content in knowledge base.

- ***my_page_thanks_count*** - variable that stores counts of all actions in

Table 5.7: Event Counts of Independent Measures corresponding to 99th percentile

	Guides	**MO**
my_page_visits_count	595	586
my_page_comments_count	22	4

which co-workers are thanking the pages' creators for content they had created. This is done using *Thank you* button described in Chapter 3. Hence, if an employee find any content appealing, helpful or just interesting, she can use this button to inform the creator about it - thank her for the contribution. We assume that this action positively affects creator's intention to generate more content (pages, edits or comments). As discussed in the beginning of this subsection, the data on *my_page_thanks_count* is missing for MO system. Summary statistics in Table 5.6 suggest that this tool is used the least frequently (the maximum amount of thank-you actions in a certain week is only 8) and in 95.48% is not utilized at all. Again, the measure suffers from overdispersion, which can be result of different data generating process for nulls and positive counts.

- ***KBsize*** - measures the size of knowledge base in a given one-week period by taking the sum of all pages in the end of a respective week. The number of pages in a knowledge base rises as users create new content and in both our samples, there is at least one new page created in each of studied weeks. Because of these properties, *KBsize* variable is increasing with *week_code* and is the same within users.[7] In other words, size of knowledge base corresponding to a selected week is the same for all studied users in a data. Table 5.6 shows its minimums and maximums suggesting the need of linear transformation that would scale down the values (due to possible computational problems). In regression, we will thus, use this variable scaled down to thousands (*KBsize/1000*). We can also see that knowledge-base size of MO system was at the end of our observation periods almost two and a half times larger than Guides one - the maximum *KBsize* of Guides and MO is 10,290 and 24,075, respectively.

  On the one hand, size of knowledge base provides higher opportunity

[7]Due to the balanced panels that capture the activity of all studied users in all one-week periods, *KBsize* represents the recurring sequence of values that increase as *week_code* rises.

to share knowledge by means of commenting and editing (simply, more pages offer more space for discussion and corrections). Hence, we are expecting the positive effect of this factor on our dependent variables corresponding to these events (*edit_page_count* and *comment_page_count*). On the other hand, growing knowledge base might mean that new pages are less needed. As a result we expect this factor to have negative effect on creation of pages in a system (*create_page_count*).

- In the last column of Table 5.6 we can find "*Dummy for FE*" measure. This variable represents any of dummies we decided to incorporate to datasets as a result of panel data transformation into cross-sectional. Because, these dummies stand for individual effects in balanced panel structure, their summary statistics are the same. Thus, *Dummy for FE* represents the single generalization that is identical for any individual-effect dummy in given dataset. In case of Guides data, we analyze behavior of 15 users therefore, we will add 15 dummy variables, each equal to one if and only if the observations of dependent measures for given week relates to the user whose individual effect we are considering (in case of MO we will add 11 dummies). Because we are dealing with the data with low number of users (panel variable) and high number of one-week periods (time variable) we were able to perform the transformation. While, these variables are added to regression to specify fixed effects in the model and to treat it unconditionally (see Chapter 6) we do not study their effects on dependent measures.

5.4 Data for the Second Hypothesis

Hypothesis #2: *Further content creation (creation of pages, edits of pages and comments) is promoted by gamified tools, concretely by viewing placements in Hall of Fame page and by previous positions reached in Hall of Fame leader-board.*

We obtained the data for our second hypothesis using the same extraction procedure as in case of first hypothesis. The variable of our interest is again continuous knowledge generation defined as an activity leading to creation of pages, page edits and insertion of comments. In Section 2.3 we introduced gamification concept that can be used in knowledge management systems in order to promote activity of users. Semanta, in its knowledge bases, directly

incorporates a gamified tool represented by Hall of Fame page (Section 4.3). Since gamification is relatively new area of interest (at least in connection to knowledge bases), Semanta started to monitor activity within Hall of Fame page only five months before completion of this thesis, in February, 2015. This limited our observation period to only 15 weeks, more precisely, our observation period begins in February 16, 2015 and ends in June 5, 2015. Unfortunately, 15 weeks were not sufficient for MO system to deliver data that would have been appropriate for our study. The main reason for this is that MO system is less frequently used than Guides (discussed in the beginning of this chapter).[8] In this section we are thus considering only Guides knowledge management system for analysis.

Extracted panel data consists of 15 one-week periods for each observed participant (*week_code*). Again, we are dealing with the problem of considerable amount of zero observations for less active users. Similarly to methodology employed in Section 5.3, we include only users with at least one activity of our interest (creation of pages, commenting and editing) in two weeks on average. Because we have 15 one-week periods, all users for whom the sum of frequencies of her activity is less than 7 are excluded. The resulting number of studied users after removing those who were inactive is 13. Therefore, our panel data consists of 13 individuals over 15 time periods. We present the summary for the dataset in Table 5.8.

Table 5.8: Dataset Summary for Second Hypothesis

	Guides
No. of one-week periods	15
No. of studied users	13
No. of obs.	195

The following subsection provides descriptive statistics for our dependent measure. Although we have detected overdispersion, second problem connected to event count variables - excess zeros, is not prevailing (see Section 5.3.1). Because our observation period is short (15 weeks) we do not assume zero outcomes to be structural but rather sampling. In other words, we consider the situations (weeks) in which no content was created by given individual to be caused by other than always zero regime (see Chapter 6). Hence, we will

[8] After extracting the dataset for MO system there were only 6 users with at least one activity in two weeks on average which made estimation of our chosen model impossible.

not transform the dataset into cross-sectional data as we have done in first hypothesis and we will keep the original balanced-panel structure.

5.4.1 Dependent Measure: Further Content Creation - Second Hypothesis

In contrary to first hypothesis, we are not considering the single elements of content creation: generation of new pages, commenting them or editing them. Instead, we are analyzing the effects on all the three actions as a whole. Our dependent measure, *create_content_count*, is thus defined as a sum of all activities of a given user in a given week that generate knowledge: pages, comments and edits. In other words, this variable is simply the sum of values of *create_page_count*, *comment_page_count* and *edit_page_count* defined in Subsection 5.3.1. We decided to estimate effects on overall activity rather than on each its element separately (as in first hypothesis) because of small number of observations on each studied variable. Moreover, this analysis represents one of the first attempts to assess relationship between gamified tools and knowledge-base adoption and thus, we concentrate mainly on general effects.

create_content_count represents event count variable ranging from zero, if there is no activity at all for a given user-week combination, to some positive value otherwise. The overall summary statistics together with between and within values are shown in Table 5.9.[9]

Overall, we can see that *create_content_count* varies from 0 to 60, with distribution skewed to the left and with long right tails. Expected value is shifted to the origin that is common feature of overdispersion. Moreover, we observe variance to be number of times higher than the corresponding mean value. Average amount of all pages, comments and edits created in a week by each user (between) ranges from 0.467 to 22.533. The negative values of within minimums do not mean that an user did not create any content. Actually, within values represent deviation from each individual's average corrected for the global one (overall mean). Therefore, there is some user in Guides system for whom the maximal deviation from average content creation is 47.133 (54.492 - 7.359).

The overall frequency of zero observations in *create_content_count* variable

[9]The variable $create\_content\_count_{it}$ is decomposed into between ($\overline{create\_content\_count_i}$) that is calculated over n users and into within ($create\_content\_count_{it} - \overline{create\_content\_count_i} + \overline{\overline{create\_content\_count}}$) that is together with overall values calculated over $n \times T$ user-weeks of data (source: `www.stata.com`).

Table 5.9: Detailed Summary Statistics for *create_content_count* - Second Hypothesis

	Guides		
	Overall	*Between*	*Within*
Obs.	195	195	195
Mean	7.359		
St. Dev.	11.238	6.320	9.446
Variance	126.293		
Skewness	2.330		
Kurtosis	8.706		
Min	0	0.467	-15.174
Max	60	22.533	54.492
Zeros (%)[a2]	27.18	92.31	29.44

Notes:

[a2] Proportion of zeros in a given variable.

is 27.18 %. In addition, 92.31 % of users had ever delivered no content to Guides knowledge base. The interesting statistic is offered by within values. If it ever happened that an user did not contributed in knowledge creation then 29.44 % of her observations is zero (Table 5.9). Finally, Figure 5.3 shows line plot of our dependent variable for each user studied. It indicates large variation across individuals.

5.4.2 Independent Variables - Second Hypothesis

In this part, our core interest lies in estimated effects of gamified tools installed in knowledge management system, concretely in effects of *dViewedReached* and *dNotViewedReached*. We tried several specifications of the model, after which we decided to include also *my_page_visits_count, my_page_comments _count* and *my_page_thanks_count* variables that we already discussed in Subsection 5.3.2 and also first lag of our dependent measure *create_content_count*. Our dataset is balanced and thus we do not miss any observation (there are 195 observations for all variables). Summary statistic (Table 5.10) and detailed description of each explanatory measure follows.

- ***dViewedReached, dViewedNotReached*** - in order to analyze how a gamified tool, like Hall of Fame page, affects creation of knowledge,

Figure 5.3: Graphs of *create_content_count* by *user_name*, Guides

we will employ three dummy variables, from which two will be used in regression and one will be set to base category. We can identify them as follows:

1. *dViewedReached* - takes value one if an user in a given week visited at least once the Hall of Fame page AND in this given week she was presented in any of the four leader-boards (meaning that her activity in a previous week was sufficient to reach the Hall of Fame placement), and zero otherwise
2. *dViewedNotReached* - takes value one if an user in a given week visited at least once the Hall of Fame page AND in this given week she was NOT presented in any of the four leader-boards (meaning that her activity in a previous week was NOT sufficient to reach the Hall of Fame placement), and zero otherwise
3. *base category* - takes value one if an user in a given week did NOT visit the Hall of Fame page AND either in this given week was presented OR was not presented in any of the four leader-boards, and zero otherwise

The intuition behind these measures is that visiting Hall of Fame page

Table 5.10: Summary Statistics for Explanatory Measures - Second Hypothesis

	Obs.	Mean	St.D.	Var	Min	Max	Zeros (%)[a2]
Guides							
dViewedReached	195	0.164	0.371	0.138	0	1	83.59
between			0.176		0	0.667	100
within			0.331		-0.502	1.097	83.59
dViewedNotReached	195	0.087	0.283	0.080	0	1	91.28
between			0.129		0	0.467	100
within			0.254		-0.379	1.021	91.28
my_page_visits_count	195	59.918	141.01	19884.14	0	866	19.49
between			135.786		0	504.667	38.46
within			52.696		-203.749	421.251	50.67
my_page_comments_t	195	1.692	6.534	42.699	0	59	80
between			4.409		0	16	100
within			4.966		-14.308	44.694	80
my_page_thanks_t	195	0.251	0.762	0.581	0	6	85.13
between			0.306		0	0.933	100
within			0.703		-0.682	5.318	85.13

Notes:
[a2] Proportion of zeros in a given dependent variable.

that shows leader-boards in five categories: Contributor, Commenter, Consumer, Thanks Receiver and Thanks Giver (see Section 4.2.1), should motivate individuals to create content. Firstly, users that reached some placement and visited Hall of Fame page should be driven to maintain their positions also in next week - this can be achieved only by creating further content. Secondly, if an user was not active enough within a system and did not reached any placement but she visited the Hall of Fame page, then this user should be also motivated to create further content. In this case the driver would be willingness to overrun others and to reach any placement in a following week (implied by gamification concept, Section 2.3). And finally, we assume that not visiting the Hall of Fame page means that users do not know about their positions and in this case, the principle of gamification is not in place. Thus, we assume that both, *dViewedReached* and *dViewedNotReached*, affect our dependent variable positively and compared to base category in higher extent.

Looking at the summary statistics for Guides data in Table 5.10, we

can see that there is very high proportion of zeros in *dViewedReached* and *dViewed NotReached* - 83.59 % and 91.28 %. This means that for only 16.41 % of all user-week combinations, an individual who visited Hall of Fame actually found her name in some leading category and for 8.72 % an individual learned that she had not reached any placement. An implication is that just in 25.13 % overall observations an user visited Hall of Fame page in some week. Between values say that 100 % of users in Guides system ever had $dViewedReached = 0$ and $dViewedNotReached = 0$. Important though are between frequencies for cases $dViewedReached = 1$ and $dViewedNotReached = 1$ (Tables B.1 and B.2 in Appendix B). We can observe 84.62 % of all users ever visited Hall of Fame while reached some position and 61.54 % ever visited Hall of Fame but not succeeded to be part of that week leader-board. These are quite high numbers. Moreover, conditional on an user ever had $dViewedReached = 1$, 19.39 % of her observations have also $dViewedReached = 1$. The corresponding within value for *dViewedNotReached* is 14.17 %.

- ***my_page_visits_count, my_page_comments_count***, ***my_page_thanks_count*** - these event count variables are already defined in Section 5.3.2. In second hypothesis we are, however, working with different observation period that is a lot shorter than one utilized in first hypothesis. Overall summary statistic for all these measures from Table 5.10 are very similar to those from Table 5.6 and we thus, assume them to affect *create_content_count* positively.

Finally, we also include the first lag of dependent variable *create_content_count* into our regression. The reason is that our dummy variables *dViewedReached* and *dViewedNotReached* might be considered as a proxy for all the activity within Guides system performed in a previous week. Estimated coefficients for these measures thus, may be influenced and we would like to correct for it.

5.5 Data for the Third Hypothesis

Hypothesis #3: *Knowledge base adoption (knowledge seeking) is driven by speed of response to questions and requests and by variety and amount of these answers provided by system experts.*

To study our predictions from the third hypothesis we will employ again two datasets, one from Guides and the second from MO system. However, as we discussed in Subsection 5.2.2, the character of the dataset differs from the ones studied in the previous two sections. In our third hypothesis, we do not attempt to study the activity of each user in the system during each week of analyzed period. Instead, we are analyzing only activity (knowledge seeking) of users who asked the experts for missing knowledge (who used *Ask* button, se Chapter 3). We are measuring how often these users visit the system in seven days after their request was fulfilled and not their activity in each consecutive one-week period. Therefore, we will be dealing with cross-sections of ever asked questions in respective systems rather than with the panel data of users over time. We will proceed from the time the first question was asked in a system up until March, 2015 - "End Date".[10] The resulting number of observations and the corresponding periods are shown in Table 5.11.

Table 5.11: Datasets' Summary for Third Hypothesis

	Guides	**MO**
No. of Observations	188	135
Corresponding Period	5-Apr-2013 *to* 30-Mar-2015	17-Jan-2013 *to* 30-Mar-2015

5.5.1 Dependent Measure: Knowledge Seeking

To study the effect on system adoption in sense of knowledge seeking we will use event count variable - *visitsAfterAnswer*. This measure represents the count of all knowledge-base visits of asking user during the seven days after her question was answered.[11] Our samples, however, contain 12.77 % (Guides) and 26.67 % (MO) of questions that remain unanswered at "End Date" (the end of our observation period). For such observations, this means that there is no real time from which it is possible to count the system's visits (as variable considers the starting time for the calculation the exact timestamp of answer). To deal

[10]By "first question asked in the respective system" we mean the very first row corresponding to ask-action captured in the database table not the very first request that appeared in the system (because the activity within the system was not being captured from the beginning of its operation).

[11]The period of one week was chosen because of consistency reasons (seven days period employed also in the first two hypotheses) and because it is sufficiently long term to observe the effect of explanatory variables on the dependent measure.

with this we decided to set the answer dates of unanswered cases to *time when question was created + 50 days* (Guides) and *time when question was created + 6 months* (MO). We have chosen these periods based on maximum time at which questions in each sample are still answered. Table 5.12 shows the minimum and maximum length of time between creation of a request and its resolution in hours. We can see that Guides system provides answers at latest 24 days after the question was set or leave the request without reply (in our observation period). The corresponding span in MO system is 87 days (almost 3 months). This plus the mean values suggest that experts in MO are more passive and reply less frequently. The values: 50 days in Guides and 6 months in MO are then outcomes of our assumption that creators of unanswered questions do not expect their handling in twice the maximum time to answer in respective system. In comparison to methodology of setting *visitsAfterAnswer* of unanswered requests to *N/A* or zero values (when assuming the "End Date" is the answer date), this approach provides also the possibility to estimate the effects on user's system visits when she expects that her request would not be answered. We consider the choice of doubling the maximum time to answer as the most reasonable however, we have performed the regression analyses using different lengths of periods and the results are not sensitive to such specification (see Section 7.3).

Table 5.12: Time to Answer

	Guides	**MO**
Obs.	188	135
% of unanswered	12.77	26.67
If answered		
min	0.003	0.004
mean	26.38	294.66
max	573.76	2086.39
If unanswered		
period	50 days	6 months

The summary statistics of our dependent measure for both systems are presented in Table 5.13. Apart from the previous two hypotheses, we are not dealing with excess zero problem.[12] The proportion of zeros is in both samples rather low, 6.91 % and 14.81 %. To understand it we have divided the sample into two subsamples, one that contains observations on answered

[12]This is implied by the Vuong test that compares zero-inflated negative binomial to standard negative binomial model (see Section 7.3).

questions and the other that contains unanswered ones. Comparing Guides with MO, we can see that Guides' users turn back to system every time their questions were answered and stop visiting it in 54.17 % of cases when question remains unanswered. The situation for MO differs as not all the users returns to system when their requests are resolved. Concretely, in 3.03 % when question is actually answered, creator does not visit the system in 7 days after answer date at all. The resulting proportion of "not showing" in the system if question is unanswered is 47.22 %. The overdispersion is detected, as for both systems the variance is number of times higher than mean (in case of Guides more than 120 times and in case of MO, 69 times) (Table 5.13). The respective histograms (Figure 5.4 and Figure 5.5) imply positively skewed dependent measure. Its distribution shows acute peak around the mean and fatter tails. We are thus dealing with negative binomial distribution with nonzero alpha (Chapter 6).

Table 5.13: Detailed Summary Statistics for *visitsAfterAnswer* - Third Hypothesis

	Guides	**MO**
Obs.	188	135
Mean	123.19	93.73
St. Dev.	121.9	80.56
Variance	14860.75	6489.62
Min	0	630
Max	0	286
Zeros (%)[a3]	6.91	14.81
If answered		
min	3	0
max	630	286
Zeros (%)[b3]	0	3.03
If unanswered		
min	0	0
max	288	11
Zeros (%)[c3]	54.17	47.22

Notes:

[a3] Proportion of zeros in a given variable.

[b3] Proportion of zero visits in subsample of answered questions.

[c3] Proportion of zero visits in subsample of unanswered questions.

Figure 5.4: Histogram of *visitsAfterAnswer*, Guides

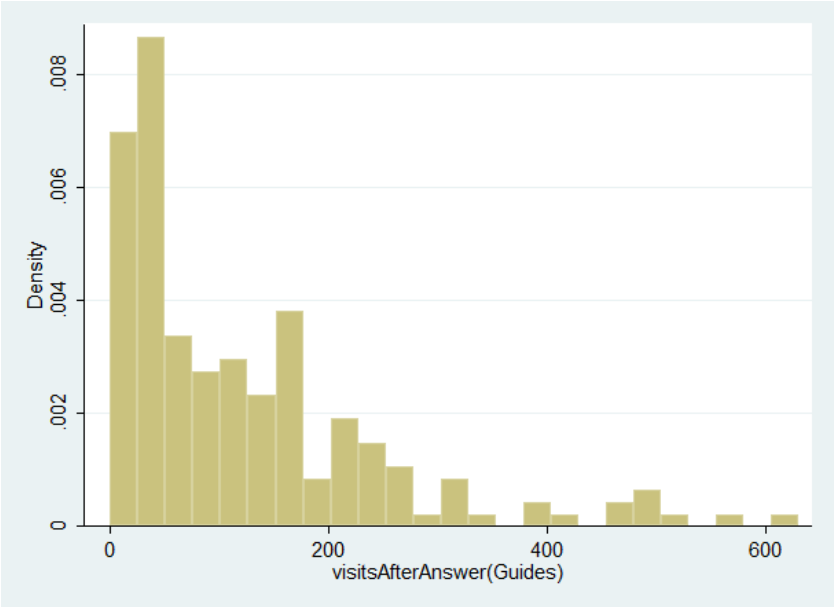
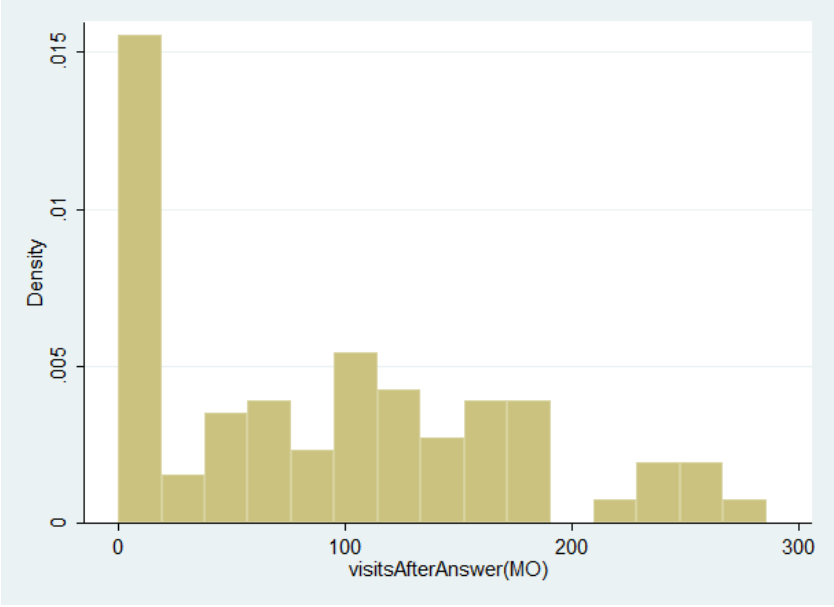


Figure 5.5: Histogram of *visitsAfterAnswer*, MO



5.5.2 Independent Variables - Third Hypothesis

To study effects on knowledge seeking induced by question mechanism (*Ask* button) we will employ these explanatory variables: *visitsBeforeQuestion, dummyHour, dummyDay, dummyWeek, numberAnswers, uniqueExperts* and *dummyMoreAnswers*. Their summary statistics are shown in Table 5.14. The detailed description follows.

Table 5.14: Summary Statistics for Explanatory Measures - Third Hypothesis

	visits Before Question	*dummy Hour*	*dummy Day*	*dummy Week*	*number Answers*	*unique Experts*	*dummy More Answers*
	Guides						
Obs.	188	188	188	188	188	188	188
Mean	163.91	0.41	0.31	0.11	2.30	1.46	0.72
Median	151.5	0	0	0	1.5	1	1
St. Dev.	90.40	0.49	0.46	0.32	3.94	1.03	0.45
Variance	8171.49	0.24	0.21	0.10	15.54	1.06	0.20
Min	1	0	0	0	0	0	0
Max	433	1	1	1	38	8	1
Zeros (%)[a3]	-	59.04	69.15	88.30	12.77	12.77	27.66
	MO						
Obs.	135	135	135	135	135	135	135
Mean	124.47	0.45	0.5	0.09	2.24	1.08	0.86
Median	118	0	0	0	2	1	1
St. Dev.	77.19	0.50	0.22	0.29	2.45	0.79	0.35
Variance	5958.56	0.25	0.05	0.08	6.02	0.63	0.12
Min	9	0	0	0	0	0	0
Max	316	1	1	1	16	3	1
Zeros (%)[a3]	-	54.81	94.81	91.11	26.67	26.67	14.07

Notes:

[a3] Proportion of zeros in a given variable.

- *visitsBeforeQuestion* - the event count variable that represents the sum of all visits to knowledge base performed by asking user in the period of seven days before the question was set. While the asking entity is the user of system who at least once visited the concrete knowledge base,

this measure is always positive (even if the question was not answered). From Table 5.14 we can see that the minimum number of system visits is 1 (Guides) and 9 (MO) what means that the asking users seek the knowledge at least once and nine times in seven days before asking for missing information. In both cases we observe overdipersion and variances considerably exceed corresponding means. We predict that if an employee seeks the knowledge before asking a question she would also look for it after her request is answered. Therefore, we suppose this variable to have nonnegative impact on our dependent measure.

- *dummyHour, dummyDay, dummyWeek* - to assess the effect of speed with which the questions are answered within a system we employ four dummy variables (three used in regression and one set as a base category) that identify the following cases:

    1. *dummyHour* - one if a question is answered within one hour, zero otherwise
    2. *dummyDay* - one if the question is answered in less than a day but more than an hour and zero otherwise
    3. *dummyWeek* - one if the question is answered in more than one day but in less than one week and zero otherwise
    4. *base category* - one if the question is answered in more than a week and zero otherwise

    This approach allows us to compare the first three cases with base category and to comment on odds of faster response to always slower answering in more than a week. Table 5.15 presents the percentage representation of each case in a respective system. The majority of questions in both, Guides and MO, are resolved in less than one hour. However, experts in MO system process the requests less frequently and users wait for their answers in 40.73 % of cases more than one week that is considerably greater proportion in comparison to Guides and its 16.49 %. From previous discussion about frequency of unanswered requests, 12.77 % (Guides) and 26.67 % (MO), we can conclude that in 3.72 % and 14.06 % users obtain answers from experts in more than one week. Following the hypothesis, we predict that faster responses positively affect users to seek knowledge and thus, we expect all three dummy variables *dummy-*

Hour, dummyDay, dummyWeek to be positive and their incidence rate ratios to be greater than one.

Table 5.15: Percentage Representation "time to answer" Cases - Third Hypothesis

	Guides	**MO**
less than 1 hour	40.96 %	45.19 %
from 1 hour to 1 day	30.85%	5.19 %
from 1 day to 1 week	11.7 %	8.89
more than 1 week	16.49 %	40.73 %

- *numberAnswers* - the event count variable representing the count of all answers provided by experts to given question. Every question in the system is in form of a page, therefore each answer appears as a comment to this page or as a single edit of the page in part "The Best Answer" (see Chapter 3). Evidently, this variable ranges from zero if request is not answered to some positive count otherwise. The proportion of zeros presented in Table 5.14 suggests that in Guides data there are 12.77% unanswered questions and in MO data 26.67% (same outcome can be seen in Table 5.12). Looking at the results in Table 5.14 we detect overdispersion and positively skewed data in both systems. The mean values indicate that there are 2 answers per question on average. *numberAnswers* measure is expected to have positive sign as we assume that the amount of answers determines the stage of interest about the question and thus, motivate the asking person to seek knowledge more (she believes in the same level of interest again).

- *uniqueExperts* - measures the number of unique experts that possess some contribution in answering. According to Table 5.14 there is one unique expert dealing with a request on average in both systems. This property together with mean outcome of *numberAnswers* suggests that on average, there is one expert who perform two replies to answer the request. Again, the proportion of zeros exactly matches the previous cases. Similarly, we predict positive relationship with dependent measure as we assume that more experts means higher involvement.

- *dummyMoreAnswers* - binary variable that takes value one if for a given user there is more than one question answered in period of 7 days during

which we measure *visitsAfterAnswer* and zero otherwise. In other words, suppose an user i asks total of N_i questions in our observation period and for $k_i = 1, ..., N_i$ she receives the answer on k_ith question in time t_{i,k_i} and $t_{i,k_i} < t_{i,k_i+1}$.[13] Then *dummyMoreAnswers* is equal to one for all k_i and k_i+1 such that $t_{i,k_i+1} - t_{i,k_i} < 7$ days. We decided to include this variable as an attempt to improve our model by controlling for observations for which the period of our interest (7 days after answer) overlaps for a given user.[14] Table 5.14 shows that there is high proportion of such events in both systems. Concretely, in 72.34 % (Guides) and 85.93 % (MO) of questions we can observe at least one other question such as their periods of 7 days after answers overlap.

[13]We do not assume equal answer dates for any pair of answers to questions asked by one user because answer dates are captured in miliseconds which makes equality impossible.

[14]We suggest more advanced statistical techniques to address this phenomenon and to correct for possible problems connected to overlapping periods of interest. However, we will not attempt to identify such methods in our analysis and recommend it for further research.

Chapter 6

Methodology

The examination of datasets in previous section showed that we are dealing with different types of data with respect to our three studied hypotheses.

In first hypothesis, the originally extracted panel suffers from two common problems connected to event counts - overdispersion and excessive proportion of zero observations. Popular statistical packages, like Stata or R, lack such mechanisms that are able to process panel count data while accounting for these two undesirable properties (at least they are not implemented yet). Thus, we decided to follow Allison & Waterman (2002) who in their paper suggest utilization of unconditional negative binomial model rather than the conditional one in analysis of panel count data. That is, to specify a conventional negative binomial regression model with dummy variables to estimate the fixed effects.[1] As a result, we will treat the data as cross-sectional and estimate the effects using well known model that copes with overdispersion as well as with excess zero problem - zero inflated negative binomial model (ZINB).

In second hypothesis, we also obtained longitudinal data structure, however, we do not assume excess zero problem besides overdispersion in event count dependent variable. Therefore, we decided to keep the original panel and estimate the effects using model that handles overdispersion in such specification - random effects (RE) negative binomial regression model for panel data.

Lastly, the datasets utilized in our third hypothesis represent cross-sectional observations on users' questions. Again, we are employing event count variable

[1] Allison & Waterman (2002) further state two possible problems connected with unconditional negative binomial model: incidental parameters problem (Greene 2012, p. 413) and problem with large number of dummies (that might cause the computational problems). The first one is neglected later in their paper, where they argue that there is a little evidence for incidental parameters bias under numerous model specifications. The rejection of second one follows from our data specification.

as dependent measure and as in all previous datasets we detected overdispersion. Hence, to deal with it we will use simple negative binomial model for cross-sections.

In application, the starting point for zero-inflated negative binomial model is standard negative binomial (NegBin) specification. Therefore, we will not divide the methodologies into parts that will correspond to hypotheses but we will firstly propose the short description of baseline Poisson regression model, then we will proceed by NegBin model from which we derive the ZINB specification. Finally we will introduce random effects (RE) NegBin model for panel data.

6.1 Poisson Regression and Negative Binomial Model

Cameron & Trivedi (2005) characterize the standard cross-section models for count data as a building block for the models that account for the special features. These are Poisson regression model and its extension: negative binomial model. The Poisson regression model specifies that y_i given x_i is Poisson distributed with density

$$f(y_i|x_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots \tag{6.1}$$

where μ_i is intensity or rare parameter and the standard assumption to derive Poisson regression model is then

$$E[y_i|x_i] = Var[y_i|x_i] = \mu_i = \exp(x_i'\beta) \tag{6.2}$$

or using log-linear model: $\ln \mu = x_i'\beta$. Log-likelihood function is then specified as

$$\ln L = \sum_{i=1}^{n} (y_i x_i'\beta - \exp(x_i'\beta) - \ln y_i!) \tag{6.3}$$

The main shortcoming of the model is so called *equidispersion assumption* according which variance is equal to mean. In general, this assumption is violated ($Var[y_i|x_i] \neq \mu_i$), which holds for our samples as well. To control for this overdisspersion issue, negative binomial (NegBin) model was introduced.

The basic idea is that NegBin generalize the Poisson model by introducing an individual, unobserved effect into the conditional mean (Long & Freese 2006),

$$\tilde{\mu}_i = \exp(x_i'\beta + \epsilon_i) = \exp(x_i'\beta)\exp(\epsilon_i), \tag{6.4}$$

where $\exp(\epsilon_i)$ is usually assumed to be gamma distributed with mean 1. Plugging $\tilde{\mu}_i$ into (6.1) we obtain the density for y_i as:

$$f(y_i|x_i) = \left(\frac{\Gamma(\alpha^{-1}+y_i)}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}, \tag{6.5}$$

where $\Gamma(.)$ represents the gamma integral that specializes to a factorial for an integer argument. The detailed derivation of Equation (6.5) can be found in Cameron & Trivedi (2005, p. 675). First two moments of negative binomial distribution are then

$$E[y|\mu,\alpha] = \mu \tag{6.6}$$

$$Var[y|\mu,\alpha] = \mu(1+\alpha\mu). \tag{6.7}$$

Both $\mu > 0$ and $\alpha > 0$ thus, the variance exceeds the mean and equidispersion assumption no longer holds (Long & Freese 2006). α is known as *dispersion paramater* since conditional variance of y increases with α. Moreover, for $\alpha = 0$ we obtain Poisson distribution.

The above mentioned model is one of the two familiar forms of negative binomial model, named Negbin 2 (Cameron 1999), that appears to have the flexibility necessary for providing a good fit to many types of count data.[2]

Cameron & Trivedi (2013) further argue that both Poisson and negative binomial regression model are not adequate if zero counts come from different processes as positive counts due to, for example, by never participating in the activity. Moreover, the presence of more zeros than predicted by count models, so called excess zeros problem, should be treated by modified specification of negative binomial (or Poisson) regression model.

[2]The Negbin 1 form of the model results if α is replaced with $\alpha = \gamma/\mu$ which leads to linear variance function $Var[y|\mu,\alpha] = \mu(1+\gamma)$ (Cameron & Trivedi 2005).

6.2 Zero-Inflated Negative Binomial Model

Zero-inflation model can be viewed as a specification in which the zero outcome can arise from one or two regimes. In one regime, there is an always zero outcome. In the second, the usual count data process is at work, which can produce zero or other positive outcome (Greene 2012). In our application, the zero outcome (no content created by an individual in one-week period) can arise as a consequence of the following situations:

1. The user is not at work or simply does not use the platform in a respective one-week period, thus she is not able to participate in content creation (always zero regime).

2. The user is not sufficiently motivated by others to create content (regime with negative binomial process).

The first situation generates always zeros, because even if the participant was motivated by other mechanisms she would not create content. The second situation implies the usual negative binomial process with zero or positive counts outcomes. In this regime, if the user was adequately motivated there would be no constraints to participate.

Another view suggests that the zero-inflation model is latent class model with two class probabilities, F_i an $1 - F_i$ (binary process with density $f_1(.)$) and the two above mentioned regimes, always zero and negative binomial data generating process (count density with $f_2(.)$) (Greene 2012). The density of such a process can be specified as (Cameron & Trivedi 2005)

$$g(y) = \begin{cases} f_1(0) + (1 - f_1(0))f_2(0) & \text{if } y = 0, \\ (1 - f_1(0))f_2(y) & \text{if } y \geq 1. \end{cases} \tag{6.8}$$

In our analysis, logit binary process will be used to determine the occurrence of each regime and then Negbin 2 described in previous section to examine the count process in second regime.

6.3 Random Effects Negative Binomial Model

Beginning this section, we are assuming the longitudinal nature of data and the dependent variable y_{it} varying over individuals ($i = 1, \ldots, n$) and over time

$(t = 1, \ldots, T_i)$. Cameron & Trivedi (2013) introduce individual-specific effect α_i that is multiplicative in conditional mean rather than additive (for count models restricted to be positive). Then

$$\mu_{it} \equiv E[y_{it}|x_{it}, \alpha_i] = \alpha_i \lambda_{it} = \alpha_i \exp(x'_{it}\beta), \tag{6.9}$$

where intercept is merged into α_i. Equation (6.9) can be also expressed as

$$\mu_{it} \equiv \exp(\delta_i + x'_{it}\beta), \tag{6.10}$$

where $\delta_i = \ln \alpha_i$. Unlike in linear models, estimator of β_j is not marginal effect but rather can be interpreted as semi-elasticity. Thus for one unit increase in x_{itj} we obtain proportional increase in $E[y_{it}|x_{it}, \alpha_i]$ (Greene 2012).

In random effects (RE) model we assume individual effects α_i (or δ_i) to be uncorrelated with regressors. Let density for the itth observation, conditional on both, α_i and regressors, denote $f(y_{it}|x_{it}, \alpha_i)$. Then joint density for the ith observation, conditional on regressors is

$$f(y_i|X_i) = \int_0^\infty \big[\prod_{t=1}^{T} f(y_{it}|\alpha_i, x_{it})\big] g(\alpha_i|\eta) d\alpha_i, \tag{6.11}$$

where $g(\alpha_i|\eta)$ is the specified density for α_i (Greene 2012). Hausman *et al.* (1984) introduced random effects negative binomial model by assuming y_{it} is NegBin distributed with parameters $\alpha_i\lambda_{it}$ and ϕ_i, where $\lambda_{it} = \exp(x'_{it}\beta)$. Then

$$E[y_{it}|\lambda_{it}, \alpha_i, \phi_i] = \frac{\alpha_i \lambda_{it}}{\phi_i}$$

and

$$Var[y_{it}|\lambda_{it}, \alpha_i, \phi_i] = (\alpha_i \lambda it/\phi_i) \times (1 + \alpha_i/\phi_i).$$

Closed form solution to (6.11) is obtained by further assuming that $(1+\alpha_i/\phi_i)^{-1}$ is beta-distributed random variable with parameters (r,s).

Other approaches to model random effects negative binomial model were introduced by Greene (2012), Cameron (1999) etc. However, we will consider the above framework for the further analysis.

6.4 Testing

Firstly, in case of all three hypotheses it would be of interest to test for overdispersion. In first and third one, for $\alpha = 0$ both equations (6.6) and (6.7) result in μ. Hence, we will test for overdispersion assuming $H_0 : \alpha = 0$ (Long & Freese 2006). In our second hypothesis we will employ test offered by Cameron (1999) where $H_0 : E[y_{it}] = Var[y_{it}]$ which means that NegBin model reduces to Poisson model. We use usual LR test with chi-square statistic specified as

$$\chi^2 = 2(\ln L_{NB} - \ln L_P), \tag{6.12}$$

where L_{NR} and L_P are the likelihood values from negative binomial and Poisson regression, respectively. Since there is only one constraint the degrees of freedom is one. Because count models are restricted to be non-negative, the usual significance level of the test is adjusted (Long & Freese 2006).

Secondly, in case of the first hypothesis we will also test whether there is an actual latent class regime splitting mechanism. Because the basic model of negative binomial and modified model is not nested, Greene (2012) suggests test statistic for nonnested hypothesis of model 1 versus model 2, proposed by Vuong (1989). This statistic is characterized as

$$v = \frac{\sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} m_i]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}} = \frac{\sqrt{n}\bar{m}}{s_m}.$$

where

$$m_i = \ln\big(\frac{f_1(y_i|x_i)}{f_2(y_i|x_i)}\big).$$

and $f_j(y_i|x_i)$ for $j = 1, 2$, denotes predicted probability that the random variable Y equals y_i. The null hypothesis is simply $E[m_i] = 0$ and interpretation of statistic is straightforward. Values of v larger than two favour model 1 whereas values less than two favour the opposite. In case $|v| < 2$, the test do not favor any model. The logic that stands behind the testing is the fact that zero-inflated induce overdispersion. Then, if the data are characterized by overdispersion, it is not obvious whether it should be credited to heterogeneity or to the regime splitting mechanisms (Greene 2012). Thus, we will produce estimates using both zero-inflated negative binomial regressions as well as original model of negative binomial without modification and compare these models

using described Vuong's test statistic.

Chapter 7

Results

In this chapter we will present the estimation outcomes of regressions associated to each of our hypothesis.

7.1 Results for the First Hypothesis

To estimate the effects on content creation given by our three dependent measures (*create_page_count, comment_page_count* and *edit_page_count*) we will employ zero-inflated negative binomial (ZINB) model with two submodels resulting from two different regimes for zero-count creation. In the first regime, zeros are created because users are not sufficiently motivated to participate. This situation is modeled by usual negative binomial model. In the second, "Allways zero regime", zero counts result from inability to participate (i.e. user is out of the work in a given week). The probability of this regime will be modeled by logit binary process. However, the probability of creating a content is expressed as a combination of the two models. Using zero-inflated negative binomial model, we will thus perform two sets (one for Guides and one for MO data) of three regressions (one for each dependent measure). We will regress the number of content created on an intercept, *my_page_visits_count, my_page_comments_count, my_page_thanks_count* (only in case of Guides), *KBsize/1000* and set of dummy variables representing individual effects - *Dummy for FE*. Specifically:

$$DV_i = \beta_0 + \beta_1^{DV} mpvc_i + \beta_2^{DV} mpcc_i + \beta_3^{DV} mptc_i \times 1(data = Guides) + \beta_4^{DV} (KBsize/1000)_i + \sum_{j=1}^{n-1} \gamma_j^{DV} DummyforFE_{ij} \quad (7.1)$$

where $mpvc_i$, $mpcc_i$ and $mptc_i$ are abbreviations for event-count explanatory variables, DV represents index for dependent measure regressed, and n is number of studied users in a given system. In the logit part of the model, we will use *my_page_visits_count* variable to estimate the probability of being in an "always zero regime".[1]

Table 7.1 shows our regression results. We decided to employ incidence rate ratios (IRR) instead of estimates of coefficients in negative binomial regression because it may more clearly communicate the influence of independent variables. IRR represents the change in the dependent variable in terms of a percentage increase or decrease, determined by the amount the IRR is either above or below 1 (Long & Freese 2006). It is an estimated rate ratio for a one unit increase in regression variable, given the other variables are held constant in the model. Because the dependent measure is actually a rate (rate is defined as a number of events per time, in our data per one week), the incidence rate ratio is simply the ratio at which the events occur. Therefore, it might be more comprehensively interpreted than usual regression coefficients that represents expected additive contributions to $\log(y)$ scale.[2] The resulting incidence rate ratios (in case of negative binomial part) and estimates of coefficients (in case of logit part) with respective p-values and standard errors (in brackets) of all explanatory variables except dummies for individual effects for each regression are shown in Table 7.1.[3] We present the complete outcome of regression analyses in Appendix A.

Firstly, we can conclude that all of the six models fit data significantly better than intercept-only models as proposed by likelihood ratio χ^2 (ll to ll_0) tests and respective p-values$<$0.001 (Table 7.1). This means that at least one coefficient per regression differs from zero. Secondly, according to p-values for natural logarithm of overdispersion parameter α for pairs *create_page_count* - Guides data and *comment_page_count* - MO data, we cannot reject the null of $\alpha = 0$ at 10 percent confidence level (implying possible better fit of Poisson regression due to failure to reject equidispersion within the data). However, the

[1] We attempted to estimate also models with other or more "inflate" variables. In case of regressing Guides' *create_page_count* and *edit_page_count*, and MO's *comment_page_count*, the estimation employing more than one such variable was not computationally feasible and only result utilizing *my_page_visits_count* in logit model predicting excessive zeros was statistically significant. In case of the remaining variables: *comment_page_count* (Guides), and *create_page_count* and *edit_page_count* (MO), we also decided to use *my_page_visits_count* to inflate zero counts based on AIC and BIC criteria comparison of other model's possibilities.

[2] For more see: `http://www.ats.ucla.edu/stat/stata/dae/nbreg.htm`

[3] Estimates of "Dummy for FE" are not reported because their effect on dependent measures are not aimed in our study.

Table 7.1: ZINB Model Reression Results (IRR) - impact of previous activity on content creation

	Guides			**MO**		
	create_page_count	**comment_page_count**	**edit_page_count**	**create_page_count**	**comment_page_count**	**edit_page_count**
	(1)	(2)	(3)	(4)	(5)	(6)
Main						
my_page_visits	1.003***	1.003**	1.002***	1.005***	1.003***	1.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
my_page_comments	1.015	1.133***	1.026*	1.117	1.396***	1.179*
	(0.016)	(0.019)	(0.014)	(0.101)	(0.111)	(0.118)
my_page_thanks	1.116	1.035	1.384***			
	(0.124)	(0.165)	(0.150)			
KBsize1000	0.802***	0.694***	0.793***	0.943***	0.898***	0.923***
	(0.036)	(0.039)	(0.029)	(0.009)	(0.008)	(0.010)
Inflate						
my_page_visits	-0.310***	-0.069***	-0.207***	-0.719***	-0.272***	-0.452***
	(0.100)	(0.022)	(0.073)	(0.144)	(0.092)	(0.085)
lnalpha						
_cons	0.159	0.328***	0.173***	0.761***	-0.274	1.190***
	(0.117)	(0.126)	(0.077)	(0.104)	(0.223)	(0.075)
Observations	930	930	930	1331	1331	1331
LR χ^2 (ll_0 to ll)[a1]	153.6***	321.3***	334.7***	188.3***	319.9***	174.4***
LR chibar2 (α=0)[b1]	440.8***	657.7***	3523.1***	534.1***	64.07***	7479.6***
Vuong[c1]	5.431***	2.041**	4.474***	7.359***	3.623***	9.047***

p-value - ***p<0.01, **p<0.05, *p<0.10, +p<0.15

Notes:

[a1] The Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero. It simply compares the model with the intercept-only model.

[b1] The likelihood-ratio test that is testing the zero-inflated poisson (zip) to zero-inflated negative binomial (zinb). The significant LR statistic for $\alpha = 0$ results in preference of zinb to zip.(Source: `http://www.stata.com/`)

[c1] Test that compares zero-inflated negative binomial with standard negative binomial (nb). The significant test indicates the better fit of zinb than nb (Chapet 6).

outcome of LR test of $\alpha = 0$ and respective p-values provided at the bottom of Table 7.1 clearly indicates that zero-inflated negative binomial is preferred to zero-inflated Poisson model for all six regressions. Moreover, Vuong test offered just below these results, which compares ZINB model with ordinary negative binomial promote favoring of our chosen model based on highly significant z-tests. To asses the interpretation of coefficients in comprehensive manner, we will proceed by dividing our discussion into two parts. In the first part we will present the results from ordinary negative binomial regression (note that we present incidence rate ratios rather than model estimates). In the second part,

we will introduce inflation logit model in which values of coefficients denotes odds of being in an "Allways zero regime". Then we will assess the discussion on overall effects.

7.1.1 ZINB - Negative Binomial Part

Looking at the incidence rate ratios (IRR) in negative binomial part (regime in which zero counts originates as a consequence of not sufficiently motivated users), we can see that number of significant explanatory measures vary across regressions. We can observe this variation only among different dependent variables specification and not among system selection (Guides or MO). Moreover, the signs of IRRs and their approximate magnitudes are equivalent across systems. Hence, we conclude that our results support the assumption that the first hypothesis can be applied on both small companies' and big corporations' knowledge management system.

The coefficient on first independent measure, *my_page_visits_count*, is statistically significant at one percent level for all of the performed regressions. The positive sign confirms our assumption that visiting pages created by a studied user by other co-workers promotes overall collaboration and contribution within a knowledge base in a given week. For example, 1 unit increase in number of creator's pages viewed in Guides system, holding other variables constant (*ceteris paribus*), results in increase of the expected rate of *create_page_count* by factor 1.003. In other words, each one-unit increase in *my_page_visits_count* elevates the expected rate of pages created by studied user by 0.03 % in the given week. The corresponding effect for MO system is 1.005 (Table 7.1).

The second coefficient's p-value of z statistic shows that *my_page_comments_count* is not a significant predictor (up to 15 percent level) in case of regressing *create_page_count*. This means that number of comments added to pages created by a user does not affect the intention of that user to create another page. However, the estimated effect on the rest two dependent measures is positive and significant. The highest rate response is estimated by regressing *comment_page_count* using MO system data and its expected change for a one-unit increase in *my_page_comments_count* is factor of 1.396 (36.6 %), *ceteris paribus*. The corresponding effect in Guides system is equal to 12.5 % (Table 7.1).

Variable *my_page_thanks_count* was employed only in Guides system. Results show its statistical importance only in case of *edit_page_count* regres-

sion. Therefore, thanking for creator's pages are inefficient motivational tool in boosting creation of pages and commenting pages. For edit-page action the situation is different and the expected rate of *edit_page_count* for each additional unit in number of thanks given to creator's pages in a given week is multiplied by factor 1.384 (increase by 38.4 %), holding other variables unchanged (Table 7.1).

The incidence rate ratios for the last variable of our interest, *KBsize1000*, show significant results in all six regressions. However, the rates are bellow one (the unit increase in *KBsize1000* leads to decrease in rate of dependent measure) and thus, do not correspond with our initial assumption that size of the knowledge base (given by a number of pages in the end of the given week) positively affects creation of edits and comments. The highest effect is detected when regressing *comment_page_count* (Guides) and each one-unit increase in *KBsize1000*, corresponding to 1000 new pages in respective knowledge-base system, multiplies the expected rate of *comment_page_count* by factor of 0.694 and thus, decrease it by 30.6 %, *ceteris paribus*. The possible explanation might be that although size of a knowledge base offers space for further collaboration, it simultaneously fills gap of knowledge required. Therefore, users do not need to create more comments when certain knowledge is already part of a system.

7.1.2 ZINB - Logit Part

The second part of the regression employs logit model to estimate the probability of being in "Always zero regime" (users do not create content because they do not have access to system, e.g. they are out of the work) relative to the regime in which knowledge is not created because users are not sufficiently motivated, although enabled to use the system. In all six regressions, *my_page_visits_count* was applied as a single inflation variable. Its coefficient is negative and highly significant (at one percent significance level) in all studied cases. We can observe stronger effects in MO regressions than in corresponding ones utilizing Guides system. This can be a result of higher proportion of zeros in MO data. Table 7.1 shows that odds of being in an excessive zero regime (always zero group) would decrease by 0.719 for every additional visit of creator's pages held by MO system while for every additional visit of creator's pages in Guides by 0.310.[4] In other words, with the increasing volume of creator's page visits by other system users, the creator's zero producing of pages (zero values

[4]The interpretation of coefficients follows from logit specification

of *create_page_count*) would be in both systems less likely generated from the always-zero process, e.g that creator is unable to use the respective knowledge base (MO or Guides) and more likely generated by the regime in which page creator do not produce pages because he is not motivated or because of similar reasons. Analogously, this holds for the rest two dependent measures we work with.

7.1.3 Overall Effect on Content Creation

From the regression results, we can conclude that the first three explanatory variables demonstrate nonnegative effects on content creation given by complete set of *create_page_count, comment_page_count* and *edit_page_count*. The effect of the fourth one, *KBsize1000*, is negative in all studied cases, therefore, knowledge-base size negatively affects overall content creation. Finally, *my_page_visits_count* decreases the probability of "always zero regime" on the whole in which no content is created (e.g. users do not have access to knowledge base and as a result they do not create any content).

7.2 Results for the Second Hypothesis

We will estimate effects on *create_content_count* employing first lag of our dependent measure - *L1.create_content_count*, *dViewedReached, sViewedNotReached, my_page_visits_count, my_page_comments_count* and *my_page_thanks_count* as explanatory variables.[5] Our regression equation is specified as:

$$\begin{aligned} create\_content\_count_{it} = \beta_0 + \beta_1 L1.create\_content\_count_{it} + \beta_2 dViewedReached_{it} \\ + \beta_3 dViewedNotReached_{it} + \beta_4 my\_page\_visits\_count_{it} \\ + \beta_5 my\_page\_comments\_count_{it} + \beta_6 my\_page\_thanks\_count_{it} \end{aligned} \tag{7.2}$$

where i is entity index (*user_name*), t is time index (*week_code*) and $L1.$ stands for first-lag operator.

We will perform the analysis using random effects (RE) negative binomial

[5] We estimated also other models with more variables as well as with different ones. Based on Likelihood-ratio test, we chose this model as the most appropriate.

regression for panel data.[6] We estimated the model also using fixed effects (FE) specification, but based on Hausman test and negative χ^2 statistic we obtained a strong evidence for not rejecting the null.[7] Thus, RE is more appropriate for our model than FE. The resulting incidence rate ratios with respective p-values and standard errors are shown in Table 7.2.

Table 7.2: NegBin RE Model Reression Results (IRR) - gamification in content creation

	Guides
	create_content_count
	(1)
Main	
L1.create_content_count	1.019**
	(0.007)
dViewedReached	1.362+
	(0.268)
dViewedNotReached	1.525
	(0.498)
my_page_visits_count	1.001
	(0.001)
my_page_comments_count	1.026*
	(0.015)
my_page_thanks_count	1.089
	(0.102)
Observations	182
LR χ^2 (ll_0 to ll)[a2]	28.29***
LR chibar2[b2]	8.27**

p-value - ***p<0.01, **p<0.05, *p<0.10, +p<0.15

Notes:

[a2] The Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero. It simply compares the model with the intercept-only model.

[b2] The likelihood-ratio test that is testing the current panel model with the pooled model (that is, a negative binomial with constant dispersion). (Source: http://www.stata.com/)

Firstly, after including first lag of dependent variable we lost 13 observations (one for each individual). The total number of observations used in our analyses

[6] We tested the model for overdispersion using Likelihood-ratio test offered by Cameron (1999) that compares log-likelihood from Poisson model with log-likelihood from negative binomial model. Null hypothesis is based on equality of conditional mean and variance, that imply no overdispersion. The resulting statistic LR=805.32814 immediately led to rejection of the null (it follows χ^2 distribution with one degree of freedom as we have only one restriction).

[7] Negative χ^2 statistic is not unusual outcome for not rejecting the null hypothesis of Hausman test in *Stata* mainly for such a small samples (source: `www.stata.com`).

is 182. Secondly, LR χ^2 statistics presented in the bottom of Table 7.2 strongly suggest that our model is statistically significant. The last important statistic offered in regression output - LR chibar2, that follows from Likelihood-ratio test of our panel model versus pooled model, suggest that negative binomial random effects specification is better than pooled one (that do not assume individual heterogeneity).

- The incidence rate ratio (IRR) for the first measure that represents our lagged dependent variable is significant at 5 % level and higher than one. This implies that, each one unit increase in user's content creation (creation of pages, comments or edits) in previous week results in multiplication of expected rate of content creation in current week by factor of 1.019, holding all other variables constant. This outcome means that if an user was active in some week, she will be active and create knowledge also in a following week.

- The core part in our analysis is estimation of effects of Hall of Fame page, that represents gamified tool in both studied systems (see Subsection 4.2.1). Dummy variables, *dViewedReached* and *dViewedNotReached*, hence, capture functionality of such defined gamification concept (see Section 2.3). The direction and magnitudes of IRRs for these measures support our assumptions that visiting Hall of Fame page (regardless of whether an user reached some placement or not) rather than not visiting it (again, regardless the positioning in leader-boards), affects content creation positively. However, only *dViewedReached* is significant at 15 % confidence level. Therefore, we can conclude that viewing Hall of Fame page and reaching the position in any monitored category (Contributor, Commenter, Consumer, Thanks Receiver and Thanks Giver) in a given week rather than not visiting it, results in elevation of content creation in that week by factor of 1.362, *ceteris paribus*. Unfortunately, we are not able to deliver any conclusions for *dViewedReached* as it may not be significantly different from our base category (not visiting Hall of Fame page). Thus, gamified tool installed in our knowledge management system induce creation of further content for those individuals who take part in gamification and who already achieved some Hall of Fame placement.

- From last three estimates in Table 7.2 only *my_page_comments_count* shows significant effect. Its direction coincides with our assumption that

comments to pages created by an user motivate her to create more content. More precisely, expected change in *create_content_count* for every additional comment to user's page is factor of 1.026 (2.6 %) in Guides system, *ceteris paribus*.

7.3 Results for the Third Hypothesis

To assess the estimated effects on knowledge seeking we will firstly regress *visitsAfterAnswer* on *visitsBeforeQuestion, dummyHour, dummyDay, dummyWeek, numberAnswers* and *uniqueExperts* using negative binomial regression model. Then we will add *dummyMoreAnswers* and check whether it significantly improves the fit of the model. This measure will be included as an attempt to control for overlapping periods of interest in our dependent variable as described in Subsection 5.5.2. We can specify the regression equation as follows:

$$\begin{aligned} visitsAfterAnswer_i = \beta_0 + \beta_1 visitsBeforeQuestion_i + \beta_2 dummyHour_i \\ + \beta_3 dummyDay_i + \beta_4 dummyWeek_i + \beta_5 numberAnswers_i \\ + \beta_6 uniqueExperts_i \end{aligned} \tag{7.3}$$

where i is cross-sectional index for question asked.

The results of the analysis (incidence rate ratios, p-values and standard errors) are shown in Table 7.3. As discusses in Section 5.5.1, we have obtained the answer dates for unanswered questions by adding double the maximum time between creating the request and answering it for a given system to the respective ask-dates (50 days for Guides and 6 months for MO). To show that resulting effects are not sensitive to such specification, we have performed also analyses using other time intervals: 30, 100 and 150 days in case of Guides, and 3, 9 and 12 month in case of MO. Resulting effects as for direction, magnitudes and significance appeared similar. The complete set of outcomes can be found in Appendix C.

Firstly, both models are statistically significant as indicated by LR chi-square tests that compare log-likelihoods of full models to intercept-only models (LR χ^2 (ll_0 to ll) statistics in the bottom part of Table 7.3). The resulting pseudo R-squared values are 0.0214 (Guides) and 0.0582 (MO). Secondly, neg-

Table 7.3: NegBin Model Reression Results (IRR) - impact on knowledge seeking

	Guides	**MO**
	visitsAfterAnswer	**visitsAfterAnswer**
	(1)	(2)
Main		
visitsBeforeQuestion	1.004***	1.009***
	(0.001)	(0.001)
dummyHour	2.015***	1.909**
	(0.461)	(0.486)
dummyDay	2.117***	1.845+
	(0.520)	(0.814)
dummyWeek	1.957**	4.722***
	(0.589)	(0.689)
numberAnswers	0.994	1.058
	(0.025)	(0.058)
uniqueExperts	1.274**	2.871***
	(0.149)	(0.648)
lnalpha		
_cons	-0.015	0.035
	(0.101)	(0.131)
Observations	188	135
pseudo R^2 [a3]	0.0214	0.0582
LR χ^2 (ll_0 to ll)[b3]	46.69***	84.98***
LR chibar2 (α=0)[c3]	1.3e+04***	5775.98***

p-value - ***p<0.01, **p<0.05, *p<0.10, +p<0.15

Notes:

[a3] Pseudo (Mc Fadden's) R-squared that measures the improvement of the fitted model to the intercept-only model.

[b3] The Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero. It simply compares the model with the intercept-only model.

[c3] The likelihood-ratio test that is testing the Poisson model to negative binomial model specification. The significant LR statistic for $\alpha = 0$ results in preference negative binomial to Poisson model or simply that the response variable is over-dispersed and is not sufficiently described by the simpler Poisson distribution (Source: http://www.stata.com/)

ative binomial specification provides better fit to both datasets than Poisson specification as suggested by highly significant ($p < 0.001$) values of chi-square test of $\alpha = 0$ (from regression output for $log(\alpha)$ we can also conclude that αs are different from zero). Finally, the regression results for both Guides and MO are very similar as for magnitude and significance of incidence rate ratios across systems (except for *numberAnswers* for which effects are insignificant and direction of impact is opposite). This implies that knowledge-base adoption followed by our third hypothesis can be applied on both small and

big companies' environments. The interpretation of results follows (again note that we employed incidence rate ratios instead of estimates of coefficients as discussed in Section 7.1).

- The first explanatory variable, *visitsBeforeQuestion*, shows highly significant positive effect on knowledge seeking in both systems. Each one unit increase in *visitsBeforeQuestion* (number of user's visits of the system in seven days before asking a question) multiplies the expected rate of *visitsAfterAnswer* by factor of 1.004 (Guides) and 1.009 (MO) holding other variables constant (Table 7.3). This means that if an user visits the respective system one time more in seven days before she asks a question then she is expected to visit the Guides system by 0.4 % more in seven days after her question was answered.

- The speed of answering is captured by our three dummy measures, *dummyHour, dummyDay* and *dummyWeek* (while the base category is set to situation when questions are answered in more than a week). For both systems, all incidence rate ratios (IRR) are higher than one which indicates that odds for all three cases (answering in less than one hour, between one hour and one day, and between one day and one week) is positive compared to our base category. For example, holding other variables constant, when questions are answered in less than one hour rather than in more than one week, expected rate of *visitsAfterAnswer* rise by factor 2.015 in Guides or 1.909 in MO. Interestingly, the IRRs' magnitudes of corresponding variables across systems do not match. In case of Guides, the highest effect among our dummies is detected in *dummyDay* and the smallest in *dummyWeek* variable. Therefore, answering between one hour and one day rather than in more than one week implies higher rise in expected rate of *visitsAfterAnswer*, *ceteris paribus*, than case of answering within one hour rather than our base, or case of answering between one day and one week rather than base (Table 7.3). The situation in MO is slightly different. The highest incidence rate ratio (IRR) is related to *dummyWeek* and the smallest to *dummyDay* measure. Moreover, the IRR of *dummyWeek* is more than two times larger than IRRs of other two measures. Answering questions between one day and one week rather than more than a week results in elevation of expected rate of MO's visits in 7 days after answer-date by factor 4.722, *ceteris paribus*. Finally, all three cases identified by our dummy variables are preferred to

answering the requests in more than one week. Hence, faster responses positively affect users' further knowledge seeking in both systems. This is an important result for "managers" of knowledge base, as they can directly affect users adoption through properly allocated system experts.

- The fifth coefficient's p-value of z statistic shows that *numberAnswers* is not significant regressor in the analysis of both Guides and MO data. This means that the amount of replies to user's request do not play a role in the user's decision to seek knowledge in the week following the answer.

- Our last variable offers interesting result in sense of comparison between two systems. For both, Guides and MO, the impact of *uniqueExperts* on user's knowledge seeking is positive and significant up to 5 percent level. However, the estimated effect in MO system is two times larger. More accurately, holding all other variables constant, each additional unique expert dealing with the user's question cause the rise in the expected rate of *visitsAfterAnswer* by factor 1.274 in Guides system while in MO the corresponding factor is 2.871 (Table 7.3). This outcome suggests that variety of answering experts improves users' knowledge seeking in knowledge bases of big corporations (MO) more than in small companies (Guides).

Now we will add *dummyMoreAnswers* variable into regression and examine if this measure significantly improves the model over the original model. To do this we will use likelihood ratio chi-square test. We will assess deviances resulting from the second model containing extra variable *dummyMoreAnswers* (M2) and from the original model (M1), and take their differences.[8] Because we add one variable, we will compare the calculated differences with chi-squared distribution with one degree of freedom. The resulting goodness of fit statistics are presented in Table 7.4. P-values in section Chi-square clearly indicate that we are not able to reject the null hypothesis that extra variable is useless at 5 % level. This is strong evidence of original model being better than one containing *dummyMoreAnswers*. We are presenting the regression outcome for the second model in Appendix C. AIC and BIC information criteria also favor

[8]Deviance is defined as two times the difference between the maximum achievable log-likelihood (each user's response serves as a unique estimate of the negative binomial parameter) and the fitted log likelihood. For deeper assessment see p.149 in Cameron & Trivedi (2005). We will use Stata's built-in command *fitstat* to obtain this measure.

alternative rather than null, however, they are not suitable for comparison as original model is nested in second model.

Table 7.4: Original Model vs. Model Including *dummyMoreAnswers* - Goodnes of Fit Statistic

	Guides			**MO**		
	M2	**M1**	**Diff**	**M2**	**M1**	**Diff**
Log-Likelihood						
full model	-1067.782	-1067.901	0.119	-687.668	-687.970	0.303
intercept-only model	-1091.244	-1091.244	0.000	-730.460	-730.460	0.000
Chi-square						
Deviance	2135.564	2135.801	-0.237	1375.335	1375.941	-0.606
LR	46.925	46.687	0.237	85.584	84.979	0.606
p-value	0.000	0.000	0.626	0.000	0.000	0.436
IC						
AIC	2153.564	2151.801	1.763	1393.335	1391.941	1.394
BIC	2182.692	2177.693	4.999	1419.483	1415.183	4.300

Chapter 8

Conclusion

This thesis analyzes knowledge-base (KB) adoption assuming intra-company interactions among workers to be its main factors. We employed data provided by Semanta, s.r.o., a company that develops and deploys knowledge management systems (KMSs) all over the world. To control for possible selection bias, we decided to include two systems for our study that differ in size and culture of a firm in which they are operating. The obtained datasets capture every activity of all system's users. We were thus, able to uniquely determine important success factors arising from day-to-day interactions among employees. We introduced two parts of KB adoption and studied how they are induced by chosen firm-level drivers. The first part represents generation of further content by knowledge-creators (1). We defined the content to be page, comment or page-edit created by a user. The second part is knowledge-creators' continuous seeking of information within knowledge management system (2). We set this process as the one in which users repeatedly visit the system.

We began by examining co-workers' collaborative activities as the first drivers of knowledge-base adoption. These activities include visiting, commenting or thanking for knowledge-creator's pages, and we studied how they affect knowledge-creators in producing further content in the system (1). The results showed that the studied drivers are overall significant and non-negatively affect further knowledge-creation. This means that collaboration of co-workers given by their activity towards our knowledge-creator (visiting, commenting or thanking for her pages) encourage the increase in knowledge-adoption rates. Moreover, the most important factor turned out to be interest in employee's knowledge which is determined by visiting her pages by other system's users.

Secondly, we employed newly defined concept of gamification to identify

additional drivers of content-creation within knowledge bases (1). Semanta's knowledge management system incorporates leader-board-based gamified tool in form of *Hall of Fame* page. We analyzed how viewing reached placements on a weekly basis together with situations in which users reached or not reached actual placements in a given week motivates these users to create more content in the system. Results showed that gamified tool when used (in our case when *Hall of Fame* page was visited) together with successfully achieved positions in its leader-board, positively influences further knowledge generation in comparison to situation when the tool is not utilized (*Hall of Fame* page is not viewed by a user). Estimation further suggested positive but insignificant effect in odds for combination viewed-leader-board & not-reaching-any-placement. Nevertheless, we showed that knowledge management systems may encourage their adoption directly by incorporating gamification in its processes.

To asses drivers for the second part of knowledge-base adoption - continuous knowledge seeking (2), we used element of Semanta's KMSs that allows employees to ask system-experts for missing knowledge (*Ask* button). Three important factors determining how users are supplied with answers were identified: speed in which knowledge was delivered to employee, variety of experts dealing with request and number of answers offered. Results showed that faster responses rather than those taking more than a week lead to significant elevation in worker's use of knowledge base (determined by number of visits) and hence system's adoption. Moreover, number of unique experts delivering answers also significantly promotes further knowledge seeking. However, outcome for the third factor was ambiguous. Therefore, we can conclude that the quicker a user is supplied with solution to her requests and the more experts are dealing with the answer the more she will search for information in knowledge base also in a period after answering - she will adopt the system. These results are important for managers of knowledge bases, because they are responsible for allocation of system's experts in the company and via this channel they can directly affect knowledge-base adoption.

After analysis of both knowledge management systems provided by Semanta, we can conclude that adoption of knowledge base within company's culture does not depend on the size or character of this culture. Due to relatively short observation period in study of "gamification" factors, we included only data obtained from smaller and less formal knowledge base, Guides. However, remaining two parts of our analysis showed that further content-creation and knowledge-seeking is driven by the same forces in both KMSs.

Overall, we have identified three areas of drivers for knowledge-base adoption. Those were collaborative activity of other system users, use of gamified tools in KMS and proper allocation of system experts available for employees' requests. We showed that these factors based on intra-company interactions among system's users, determine and significantly affect an increase of knowledge-base adoption.

We acknowledge that this thesis represents an attempt to fill the literature gap on knowledge-base adoption. We offer an innovative approach in determining important success factors for system's adoption by modeling relationships between its two studied parts (continuous creation of content and continuous seeking for knowledge) and intra-company interactions among employees. As far as we know, this is also the first paper that studies effects of the gamified tools in area of knowledge management.

Finally, we raise some recommendations for further research. Another analysis should process excessive zero observations in dependent variables in study of effects on content creation using zero-inflated negative binomial model for panel data. This methodology was not implemented in statistical packages when this thesis was completed. Next, due to very short observation period for our "gamification" study, we were not able to deliver robust conclusions for one of the studied knowledge management systems. We thus recommend longer period of observations entering the analysis. Lastly, the examination of continuous knowledge-seeking do not control for observations for which there was more than one request answered in same time. We thus suggest more advanced statistical techniques in working with data to address this phenomenon.

Bibliography

Aaltonen, A. & S. Seiler (2014): "Wikipedia : the value of open content production." *CentrePiece* pp. 11–13.

Abril, R. M. (2007): "The Dissemination and Adoption of Knowledge Management Practices Behavioural Model." **5(2)**: pp. 131–142.

Allison, P. D. & R. P. Waterman (2002): "Fixed-Effects Negative Binomial Regression Models." *Sociological Methodology* **32(1)**: pp. 247–265.

Arkes, H. R. & C. Blumer (1985): "The Psychology of Sunk Cost." *Organizational Behavior and Human Decision Processes* **35**: pp. 124–140.

Ballance, C. (2013): "Strategic Ways to Develop Game- Based Learning for High ROI." *T+D* **(September)**: pp. 76–78.

Barberis, N. C. (2013): "Thirty Years of Prospect Theory in Economics: A Review and Assessment." *Journal of Economic Perspectives* **27(1)**: pp. 173–196.

Cameron, A. C. (1999): "Essentials of Count Data Regression." *Cambridge University Press* pp. 1–17.

Cameron, A. C. & P. K. Trivedi (2005): *Microeconometrics: Methods and Application*. New York: Cambridge University Press, 1st edition.

Cameron, A. C. & P. K. Trivedi (2013): "Count Panel Data." .

Davenport, B. T. H., L. Prusak, & A. Webber (1989): "Working Knowledge : How Organizations Manage What They Know." .

Davis, F. D. (1992): "Perceived Usefulness, Percieved Ease of Use, and User Acceptance of Information Technology." *MIS Quarterly* **16(2)**.

Deterding, S. & D. Dixon (2011): "From Game Design Elements to Gamefulness : Defining â€śGamification â€ť." In "Proceedings from MindTrek '11. Tampere, Finland: ACM," .

Farzan, R. & P. Brusilovsky (2011): "Encouraging user participation in a course recommender system: An impact on user behavior." *Computers in Human Behavior* **27(1)**: pp. 276–284.

Farzan, R. & J. DiMicco (2008): "When the experiment is over: Deploying an incentive system to all the users." *Symposium on Persuasive Technology* .

Farzan, R., J. DiMicco, & D. Millen (2008): "Results from deploying a participation incentive mechanism within the enterprise." In "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems," .

Ghosh, D. & A. Vogt (2012): "Outliers : An Evaluation of Methodologies." *JSM* pp. 3455–3460.

Gjurovikj, A. B. (2000): "Knowledge Management as a Competitive Advantage of Contemporary Companies." In "Proceedings of the International Conference in Intellectual Capital, Knowledge Management & Organizational Learning," pp. 482–489.

Gopalakrishnan, S. & P. Bierly (2001): "Analyzing innovation adoption using a knowledge-based approach." *Journal of Engineering and Technology Management* **18(2)**: pp. 107–130.

Greene, W. (2012): *Econometric analysis*. 7. Upper Saddle River, 7 edition.

Hamari, J. (2011): "Perspectives from behavioral economics to analyzing game design patterns : loss aversion in social games."

Hamari, J. (2013): "Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service." *Electronic Commerce Research and Applications* **12(4)**: pp. 236–245.

Hamari, J. & H. Sarsa (2014): "Does Gamification Work ? - A Literature Review of Empirical Studies on Gamification." In "proceedings of the 47th Hawaii International Conference on System Science," Hawaii, USA.

Hausman, J., B. H. Hall, & Z. Griliches (1984): "HausmanHallGriliches E84.pdf." *Econometrica* **52(4)**: pp. 909–938.

Hou, C.-K. (2014): "User Acceptance of Business Inteligence Systems in Taiwan's Electronic Industry." *Social Behavior and Personality* **42(949)**: pp. 583–596.

Huang, L.-s. & C.-p. Lai (2014): "Knowledge management adoption and diffusion using structural equation modeling." *Global Journal of Business Research* **8(1)**: pp. 39–57.

Kahneman, D. & A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* **47(2)**: pp. 263–292.

Kaula, R. (2015): "Business Intelligence Rationalization : A Business Rules Approach." *International Journal of Information, Business and Management* **7(1)**.

Kuo, R.-Z. & G.-G. Lee (2011): "Knowledge management system adoption: exploring the effects of empowering leadership, task-technology fit and compatibility." *Behaviour & Information Technology* **30(1)**: pp. 113–129.

Leeson, B. C. (2013): "Driving KM behaviors and adoption through." *KM World* **(April)**: pp. 10–20.

Long, J. & J. Freese (2006): *Regression models for categorical dependent variables using Stata*. Stata Press, rev. edition.

Moise, D. (2013): "Gamifieation - The New Game in Marketing."

Nicholson, S. (2012): "A User-Centered Theoretical Framework for Meaningful Gamification." In "Games+Learning+Society 8.0," Madison, WI.

Ofek, E. & M. Sarvary (2001): "Leveraging the Customer Base : Creating Competitive Advantage Through Knowledge Management." *Management Science* **47(11)**: pp. 1441–1456.

Pink, D. H. (2009): *Drive, the Surprising Truth about What Motivates US*. New York: RIVERHEAD BOOKS.

Ritchie, W. J., S. a. Drew, M. Srite, P. Andrews, & J. E. Carter (2011): "Application of a Learning Management System for Knowledge Management:

Adoption and Cross-cultural Factors." *Knowledge and Process Management* **18(2)**: pp. 75–84.

Smith, R. G. (1985): "Knowledge-Based Systems." In "Proceedings of Canadian High Technology Show. Landowne Park, Ottawa.", .

Suresh, A. (2013): "Knowledge Management Adoption, Practice and Innovation in the Indian Organizational Set Up: An Empirical Study." *Journal of IT and Economic Developement* **4(2)**: pp. 31–42.

Vuong, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* **57(2)**: pp. 307–333.

Wong, K. Y. & E. Aspinwall (2005): "An empirical study of the important factors for knowledge-management adoption in the SME sector." *Journal of Knowledge Management* **9(3)**: pp. 64–82.

Yeoh, W. & A. Koronios (2010): "Critical Success Factors for Business Intelligence Systems." *Journal of Computer Information Systems* pp. 23–32.

Appendix A

Figure A.1: Histogram of *comment_page_count*, Guides

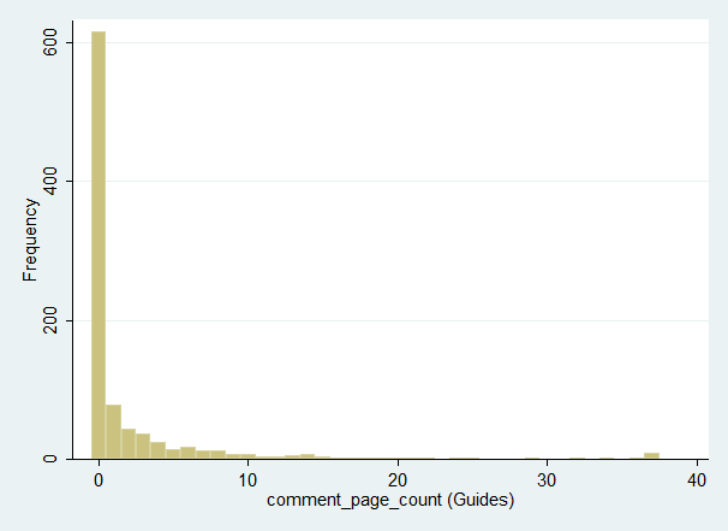


Figure A.2: Histogram of *edit_page_count*, Guides

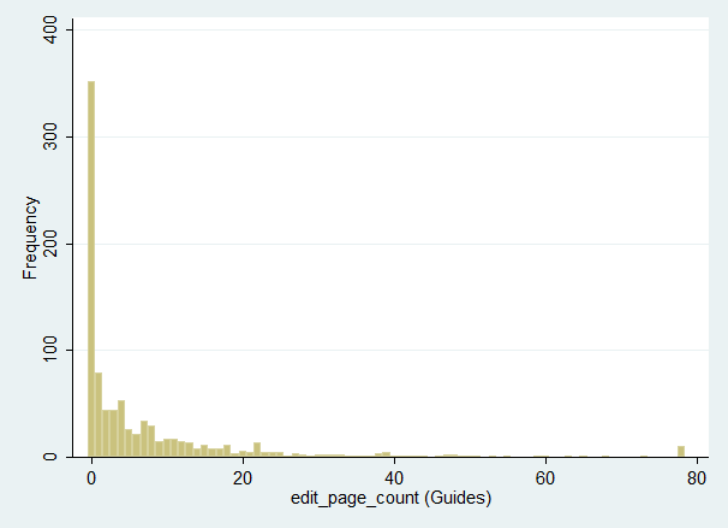


Figure A.3: Histogram of *comment_page_count*, MO

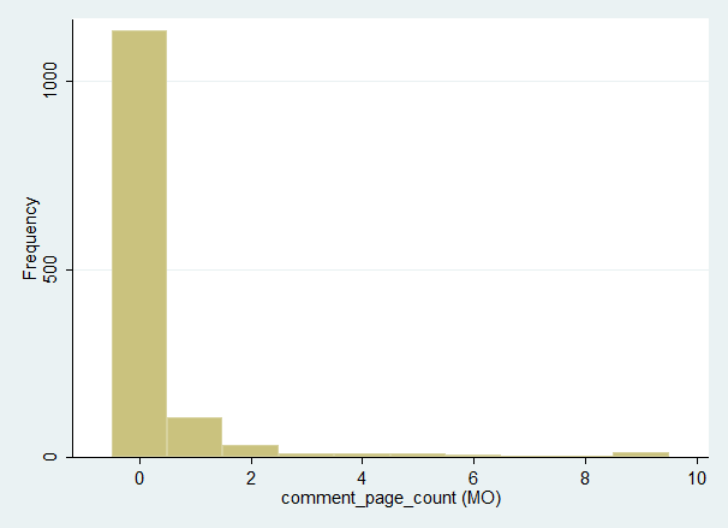


Figure A.4: Histogram of *edit_page_count*, MO

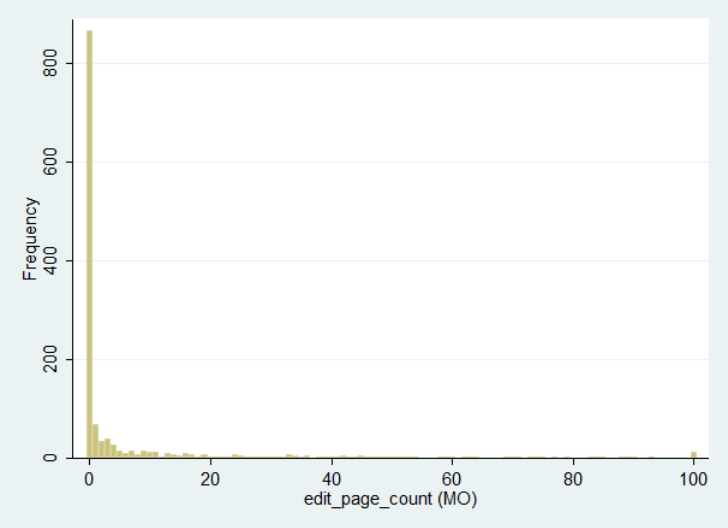


Table A.1: Complete ZINB Model Reression Results (IRR) - impact of previous activity on content creation, Guides

	Guides		
	create_page_count	**comment_page_count**	**edit_page_count**
	(1)	(2)	(3)
Main			
my_page_visits	1.003***	1.003**	1.002***
	(0.001)	(0.001)	(0.001)
my_page_comments	1.015	1.133***	1.026*
	(0.016)	(0.019)	(0.014)
my_page_thanks	1.116	1.035	1.384***
	(0.124)	(0.165)	(0.150)
KBsize1000	0.802***	0.694***	0.793***
	(0.036)	(0.039)	(0.029)
user1	1.839+	0.030***	0.327***
	(0.768)	(0.010)	(0.070)
user2	0.230**	0.037***	0.060***
	(0.146)	(0.014)	(0.016)
user3	1.115	0.034***	0.264***
	(0.616)	(0.017)	(0.090)
user4	0.674	0.088***	0.100***
	(0.290)	(0.024)	(0.022)
user5	1.671	0.129***	0.146***
	(0.683)	(0.034)	(0.032)
user6	1.342	0.088***	0.413***
	(0.567)	(0.027)	(0.089)
user7	2.446**	0.073***	0.236***
	(0.993)	(0.020)	(0.052)
user8	6.575***	0.022***	0.016***
	(2.973)	(0.009)	(0.005)
user9	1.367	0.099***	0.244***
	(0.568)	(0.027)	(0.053)
user10	1.225	0.005***	0.073***
	(0.655)	(0.004)	(0.020)
user11	0.795	0.043***	0.074***
	(0.460)	(0.017)	(0.232)
user12	4.857***	0.020***	0.232***
	(2.380)	(0.010)	(0.062)
user13	2.276**	0.083***	0.168***
	(0.924)	(0.023)	(0.037)
user14	1.360	0.037***	0.150***
	(0.673)	(0.016)	(0.041)
Intercept	3.916***	212.591***	174.185***
	(2.000)	(97.451)	(56.249)
Inflate			
my_page_visits	-0.310***	-0.069***	-0.207***
	(0.100)	(0.022)	(0.073)
lnalpha			
_cons	0.159	0.328***	0.173***
	(0.117)	(0.126)	(0.077)
Observations	930	930	930

p-value - ***p<0.01, **p<0.05, *p<0.10, +p<0.15

Table A.2: Complete ZINB Model Reression Results (IRR) - impact of previous activity on content creation, MO

	MO		
	create_page_count	**comment_page_count**	**edit_page_count**
	(1)	(2)	(3)
Main			
my_page_visits	1.005***	1.003***	1.006***
	(0.001)	(0.001)	(0.001)
my_page_comments	1.117	1.396***	1.179*
	(0.101)	(0.111)	(0.118)
KBsize1000	0.943***	0.898***	0.923***
	(0.009)	(0.008)	(0.010)
user1	0.391**	1.623	2.831***
	(0.147)	(0.781)	(0.972)
user2	0.459*	4.552***	0.489*
	(0.215)	(2.302)	(0.200)
user3	0.726	4.104***	3.447***
	(0.291)	(1.985)	(1.324)
user4	3.226***	18.755***	5.970***
	(0.962)	(6.889)	(1.832)
user5	4.370***	1.304	4.215***
	(1.970)	(1.192)	(2.012)
user6	14.781***	2.624+	19.352***
	(5.411)	(1.744)	(7.521)
user7	3.122***	1.650	3.580***
	(1.107)	(0.920)	(1.328)
user8	4.808***	0.908	2.164*
	(1.882)	(0.785)	(0.961)
user9	2.136**	16.993***	3.800***
	(0.812)	(7.893)	(1.471)
user10	2.096**	13.412***	4.630***
	(0.672)	(5.386)	(1.506)
Intercept	3.916***	0.276***	6.274***
	(2.000)	(0.117)	(2.212)
Inflate			
my_page_visits	-0.719***	-0.272***	-0.452***
	(0.144)	(0.092)	(0.085)
lnalpha			
_cons	0.761***	-0.274	1.190***
	(0.104)	(0.223)	(0.075)
Observations	1331	1331	1331

p-value - ***p<0.01, **p<0.05, *p<0.10, +p<0.15

Appendix B

Table B.1: Decomposition of *dViewedReached* Counts into Between and Within Values - Second Hypothesis

	Overall		**Between**		**Within**
	Freq.	*Percent*	*Freq*	*Percent*	*Percent*
Guides					
0	163	83.59	13	100	83.59
1	32	16.41	11	84.62	19.39
Total	195	100	24	184.62	54.17
MO					
0	163	83.59	13	100	83.59
1	32	16.41	11	84.62	19.39
Total	195	100	24	184.62	54.17

Table B.2: Decomposition of *dViewedNotReached* Counts into Between and Within Values - Second Hypothesis

	Overall		**Between**		**Within**
	Freq.	*Percent*	*Freq*	*Percent*	*Percent*
Guides					
0	178	91.28	13	100	91.28
1	17	8.72	8	61.54	14.17
Total	195	100	21	161.54	61.90
MO					
0	163	83.59	13	100	83.59
1	32	16.41	11	84.62	19.39
Total	195	100	24	184.62	54.17

Appendix C

Table C.1: NegBin Model Reression Results (IRR) - impact on knowledge seeking using *dummyMoreAnswers*

	Guides	**MO**
	visitsAfterAnswer	**visitsAfterAnswer**
	(1)	(2)
Main		
visitsBeforeQuestion	1.004***	1.009***
	(0.001)	(0.001)
dummyHour	2.123***	1.916**
	(0.537)	(0.493)
dummyDay	2.270***	1.641
	(0.646)	(0.764)
dummyWeek	2.097**	4.849***
	(0.699)	(1.739)
numberAnswers	0.993	1.062
	(0.024)	(0.059)
uniqueExperts	1.280**	2.865***
	(0.150)	(0.648)
dummyMoreAnswers	0.907	0.808***
	(0.183)	(0.225)
lnalpha		
_cons	-0.016	0.030
	(0.101)	(0.132)
Observations	188	135
pseudo R^2	0.0215	0.0586
LR χ^2 (ll_0 to ll)	46.92***	85.58***
LR chibar2 (α=0)	1.3e+04***	5776.25***

p-value - ***p<0.01, **p<0.05, *p<0.10, +p<0.15

Table C.2: Estimation results (NegBin) - *visitsAfterAnswer* (Guides)

	Guides			
	visitsAfter Answer	**visitsAfter Answer**	**visitsAfter Answer**	**visitsAfter Answer**
	(30 days)	(50 days)	(100 days)	(150 days)
Main				
visitsBeforeQuestion	1.005***	1.004***	1.004***	1.004***
	(0.001)	(0.001)	(0.001)	(0.001)
dummyHour	1.945***	2.015***	2.171***	2.845***
	(0.424)	(0.461)	(0.559)	(0.680)
dummyDay	2.054***	2.117***	2.281***	2.970***
	(0.481)	(0.520)	(0.630)	(0.759)
dummyWeek	1.937***	1.957***	2.126***	2.772***
	(0.554)	(0.589)	(0.721)	(0.878)
numberAnswers	0.997	0.994	0.992	0.983
	(0.023)	(0.025)	(0.027)	(0.025)
uniqueExperts	1.237**	1.274**	1.289**	1.406**
	(0.134)	(0.149)	(0.170)	(0.179)
Intercept	21.357***	21.822***	19.100***	12.701***
	(5.124)	(5.301)	(5.381)	(3.519)
lnalpha				
_cons	-0.159	-0.015	0.222	0.109
	(0.102)	(0.101)	(0.102)	(0.103)
alpha				
_cons	0.853	0.985	1.248	1.115
	(0.087)	(0.099)	(0.127)	(0.115)
Observations	188	188	188	188
Pseudo R^2	0.024	0.021	0.019	0.025
LR χ^2 (ll_0 to ll)	52.63***	46.69***	40.29***	54.40***
LR chibar2 (α=0)	1.2e04***	1.3e04***	1.4e04***	1.3e04***

p-value - ***p<0.01, **p<0.05, *p<0.10, +p<0.15

Table C.3: Estimation results (NegBin) - *visitsAfterAnswer* (MO)

	MO			
	visitsAfter Answer	**visitsAfter Answer**	**visitsAfter Answer**	**visitsAfter Answer**
	(3 months)	(6 months)	(9 months)	(12 months)
Main				
visitsBeforeQuestion	1.009***	1.009***	1.010***	1.010***
	(0.001)	(0.001)	(0.002)	(0.002)
dummyHour	1.908***	1.909**	2.024**	1.981**
	(0.473)	(0.486)	(0.562)	(0.571)
dummyDay	1.778+	1.845+	2.145+	1.969+
	(0.749)	(0.815)	(1.056)	(0.982)
dummyWeek	4.473***	4.722***	5.356***	4.895***
	(1.526)	(1.689)	(2.144)	(1.981)
numberAnswers	1.061	1.059	1.061	1.062
	(0.055)	(0.058)	(0.064)	(0.065)
uniqueExperts	2.539***	2.871***	3.270***	2.816***
	(0.537)	(0.648)	(0.823)	(0.705)
Intercept	3.948***	3.336***	2.243**	2.983***
	(1.023)	(0.919)	(0.741)	(0.977)
lnalpha				
_cons	-0.063	0.035	0.251	0.274
	(0.131)	(0.131)	(0.136)	(0.137)
alpha				
_cons	0.934	1.035	1.286	1.315
	(0.123)	(0.136)	(0.175)	(0.180)
Observations	135	135	135	135
Pseudo R^2	0.059	0.058	0.053	0.049
LR χ^2 (ll_0 to ll)	86.37***	84.98***	76.96***	70.03***
LR chibar2 (α=0)	5558.99***	5775.98***	5950.11***	5830.63***

p-value - ***p<0.01, **p<0.05, *p<0.10, +p<0.15