Prediction of User Behavior and Users Segmentation in a Government-owned Digital Application Using Markov Chain Model

Novaldy Pratama Putra[1*], Putri Wikie Novianti[1]

[1]Govtech Edu by Metra-Net (TELKOM Group), Tangerang, Indonesia

ABSTRACT

Kurikulum Merdeka (KM) was launched by the Ministry of Education, Culture, Research and Technology in Indonesia as an alternative curriculum during the pandemic era. It uses a digital application named Platform Merdeka Mengajar (PMM) for its implementation. As of August 2022, among the five main digital products in PMM, the most popular feature in PMM is Toolkit, which then can also be considered as the users' entry point of PMM. In order to approach every single segment in a tailored manner to promote the desired-outcome, the team is trying to identify PMM's user segmentation and predict future segment movement based on the current users' behavior in the platform. This paper focuses on developing a Markov Chain Model to predict user behavior and customer segmentation levels on Toolkit based on recency and frequency values of users' download activity. During the twelve months of the observational period, this study comprised more than 400 thousand unique users who downloaded teaching-materials. As a stochastic approach, Markov chain model predicts future behavior of 16.76% of users to remain using the product. Further, the prediction on users' segmentation using recency and frequency framework, yielded 27.98% of users would be classified as promising users. Such results were used by the team to formulate a better intervention strategy to improve the acquisition and retention in the PMM platform, such as by content enrichment and push notifications for updated content especially to possible promising segments.

Keywords: Customer Segmentation, Retention, Markov Chain, Kurikulum Merdeka.

1. INTRODUCTION

"Kurikulum Merdeka" (KM; Emancipated Curriculum) was launched by the Ministry of Education, Culture, Research and Technology (MoECRT) on February 11th 2022, as an alternative teaching curriculum to the 2013 existing-curriculum and the Emergency-Period Curriculum. KM gives autonomy to schools to determine which the more suitable curriculum for their needs is, which further translates into an emancipated learning (or Merdeka Belajar). It is the MoECRT's approach to allow for flexibility, localisation and individualisation given the diversity in the contextual needs across Indonesia. As of September 2022, KM is implemented in more than 142 thousands of schools or covers 67% of the total schools in Indonesia.

With the massive growth of internet-used within the last decades, the digital transformation especially in the educational sector is inevitable. It is catalyzed more by the COVID-19 pandemic, where more than fifty billion students in Indonesia immediately could not attend face-to-face learning at schools and pushed more than 4 million teachers to pivot their traditional-offline teaching process towards an online-environment. Although technological advancements offers higher accessibility and coverage through implementation of e-learning, as well as accelerate the feedback loop of policy (including new curriculum) implementation (Vavoula, 2008; Syahrul, 2022); the key challenges on the internet infrastructures and on the skills of end-users (i.e. teachers and students) should be also highly considered (Ja'ashan, 2020).

* *Corresponding Author. aldy.novaldy@gmail.com*

The launching of the aforementioned KM curriculum is accompanied by a digital platform named Platform Merdeka Mengajar (pen: PMM), as a digital tool to assist and to facilitate teachers in implementing KM. This digital platform consists of five features, namely Toolkit, Assessment, Micro learning, Career (Professional Activity Log), and Community-based learning. Based on the number of visiting users, Toolkit has received the highest traffic in PMM (Figure 1). Hence, it can be further considered as the entry point of PMM.

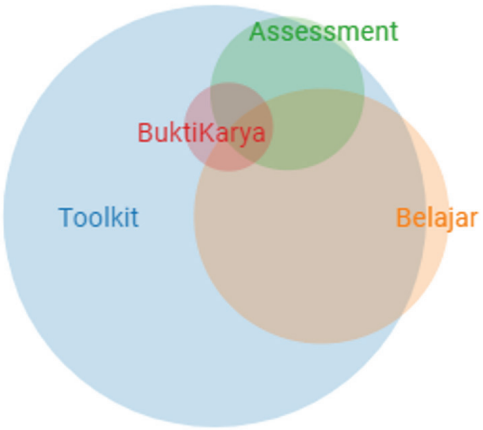

Figure 1. Venn diagram of the number of users in the five PMM's products

The team is trying to identify PMM's user segmentation and predict future segment movement based on users' behavior in the platform, in an attempt to approach each and every single segment in a tailored manner to promote the desired-outcome (i.e. activities of users to download teaching-materials). Because there is an increasing demand to more accurately forecast user future behavior, it is necessary to create a mechanism for identifying users, as the first step in developing strategies to log the interests of users and the firm (Burelli, 2019). The main objective of the current study is to create a stochastic framework to predict movement of PMM users PMM, by utilizing the big data resulting from the users' interaction in the PMM digital platform.

2. LITERATURE REVIEW

2.1 Markov Chain Model

The Markov Chain is one of many prediction techniques used in data mining. The Markov Chain only needs the most recent data to make predictions (Roman and Porto, 2008), not continuous historical data (Zhi-Hang, 2005). The Markov Chain has reasonably high accuracy, understandable findings, simple computations, and reliability based on the research that has been done (Song, 2009). The Markov Chain model has been widely applied in various studies, from (i) predicting the progression of diabetic retinopathy in type-2 diabetes patients (Srikanth, 2015), (ii) predicting the type of dangerous diseases (Mustakim and Syaifullah, 2015); to (iii) prediction of Urban Spatial Changes Pattern (AbdulRazak, 2022).

2.2 RFM-clustering

The customer management process is made using a variety of models in an effort to increase client loyalty. One of the most popular models in this regard is based on the sales metrics of recency, frequency, and monetary (RFM). Hughes introduced this approach to assess a customer's activity and to create predictions based on that behavior in marketing, (Hughes, 1994). RFM model has been widely applied in various studies, there are Estimating customer lifetime value based on RFM analysis of

customer purchase behavior (Khajvand, 2011), market segmentation using RFM model (Roshan, 2017), and implementation that combines the two methods mentioned above provides a stochastic dynamic programming model with a Markov chain that explicitly focuses on the customer, as well as a new model for valuing customers in the banking industry (Bekamiri, 2020).

3. RESEARCH METHODOLOGY

3.1 Markov Chain Discrete Model

The main idea of the Markov chain is basically a state transition, with the property that if the state of the current moment is known, the probability of the state of the process at one step forward is then only affected by the current state of the process. A Markov chain can transit and remain in the same state from time t to time t+1 (Komorowski, 2015). In other words, the state of the process in the past does not affect the future state.

Markov Chain can be applied in evaluating system reliability, if it fulfills these following requirements (Mustakim and Syaifullah, 2015):

- The system must be stationary or homogeneous, meaning that the behavior of the system is always the same at all times or the probability of a system transitioning from one state to another which will always be the same throughout time.Second point
- State is identifiable. Conditions that may occur in the system must be identified clearly,

the Markov Property in mathematical notation is formulated as follows: the Markov Property in mathematical notation is formulated as follows:

$P(X_{t+1} = s \mid X_t =, X_{t-1} = s_{t-1}, \ldots , X_0 = s_0) = P(X_{t+1} = s \mid X_t = s_t)$

for all $t = 1, 2, 3, \ldots, n$ and or for all states $s_0, s_1, s_2, \ldots, s_t, s$

Let $X_0, X_1, X_2, \ldots$ be a Markov chain with state space $S = 0, 1, 2, \ldots, N$. Given each X_t is a random variable, it has a probability distribution. We can write the probability distribution of X_t as an $Nx1$ vector.

For example, consider X_0 Let be π an $Nx1$ vector denoting the probability distribution of X_0 :

$\pi = (\pi_1 \; \pi_2 \; \vdots \; \pi_N) = (P(X_0 = 0) \; P(X_0 = 1) \; \vdots \; P(X_0 = N)$

Use the Partition Rule, conditioning on X_0 :

$P(X_1 = j) = \sum_{i=1}^{N} P(X_1 = j \mid X_0 = i) \, P(X_0 = i) = \sum_{i=1}^{N} \pi_i p_{ij} = (\pi^T P)_j$

This shows that $P(X_1 = j) = (\pi^T P)_j$ for all j.

The row vector $\pi^T P$ is therefore the probability distribution of X_1. Using the partition rule as before conditioning again on X_0:

$$P(X_2 = j) = \sum_{i=1}^{N} P(X_2 = j \mid X_0 = i) \, P(X_0 = i) = \sum_{i=1}^{N} (P^2)_{ij} \pi_i = (\pi^T P^2)_j$$

The row vector $\pi^T P^2$ is therefore the probability distribution of X_2 . Then we can conclude the probability of distribution of X_t is $X_t \sim \pi^T P^t$.

Assume a Markov chain in which the transition probabilities are neither a function of time t nor n. This defines a homogeneous Markov chain or steady state. At steady state as $n \to \infty$, the distribution vector π^T settles down to a unique value and satisfies the equation

$$\pi^T P = \pi^T \quad (1)$$

This condition happens because the distribution vector value does not vary from one time instant to another at steady state. in that case is an eigenvector for with corresponding eigenvalue $\lambda = 1$. The Markov chain has reached its steady state when the Eq.1 is satisfied. (Gebali, 2008).

3.2 RFM Model

RFM analysis is the most frequently adopted segmentation technique to quantitatively rank and group users based on the recency, frequency and monetary total of their recent transactions to identify the best users and perform targeted marketing campaigns.

RFM analysis ranks each user by considering the following factors:

- Recency. Time since last order or last engaged with the product
- Frequency. Total number of transactions or average time between transactions/engaged visits
- Monetary. Total or average transaction value.

Among the three RFM measures, recency is often regarded as the most important one. However, RFM values are inclined to be firm-specific and are based on the nature of the products (Lumsden, 2008). We visualized the RFM analysis on a 2-dimensional graph (Figure 2) including its definition (Table 1) as follows.

Figure 2. Recency and Frequency Grid visualized by CleverTap

Table 1. An example of a table.

Segment	Definition	Segment	Definition
Champions	These users are the most active users. They have the highest recency and frequency scores.	Needing Attention	These users have above average recency and frequency scores.
Loyal Users	These users have the highest frequency of use with strong recency scores.	About to Sleep	These users have below average recency and frequency scores. May slip away if not engaged with.
Potential Loyalists	These users have visited the app very recently and have the potential to become loyalists or champions.	At Risk	These users have above average frequency but low recency scores. Strong candidates to re-engage.
New Users	These users are the most recent users with low frequency scores. Strong candidates to encourage repeat use.	Cannot Lose Them	These users were active at one point in your app, but haven't been back recently. Strong candidates to re-engage.
Promising	These users have high recency scores with the potential to become high frequency users.	Hibernating	These users have the lowest recency and frequency scores. May be lost.

In order to obtain Recency and Frequency value, we need raw event data (event download logs with timestamp), from which we will derive the following unique_id (email_adress), number of downloads (Frequency, F), number of days since the user's last download (Recency, R). Having found Recency and Frequency (RF) value for each user and every month, the value is then standardized by these following Formula

$$R_{standard} = 1 - \frac{R_i - R_{min}}{R_{max} - R_{min}} \quad (2)$$

$$F_{standard} = 1 - \frac{F_i - F_{min}}{F_{max} - F_{min}} \quad (3)$$

After obtaining the standard value of RF then mix and match with condition.

4. RESULTS

4.1 Descriptive Statictics

During the observational period, we collected data resulting from the users' interaction with the PMM application. We identified there were more than 400 thousands of unique users who cumulatively downloaded 3,146 teaching-materials. It is interesting to also notice that the PMM platform, particularly the Toolkit-feature, is widely used by teachers from all age ranges, from young to senior teachers (Figure 3). The platform is also used in 8.10% of the total 3T areas.

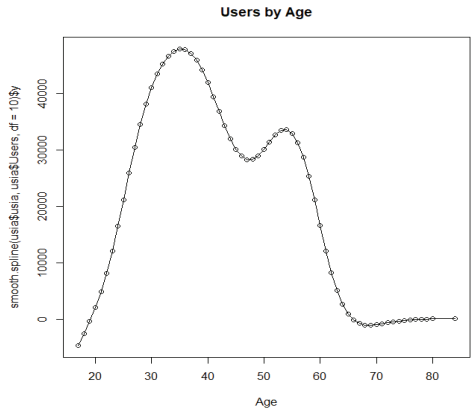

Figure 3. Total users by age

Since the application was launched, we identify an exponential growth in the cumulative number of users who downloaded teaching-materials. The number of users who downloaded teaching materials on a daily basis follows a seasonal-trend over the observational period (Figure 4). We identified the median of users' retention rate is ~2%, which then triggered the team to seek for a strategy to improve the retention of users in the platform.

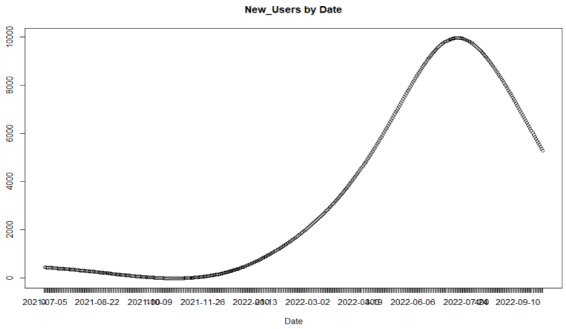

Figure 4. General pattern on the number of user download (smoothed spline)

4.2 Prediction of users' Behavior in downloading teaching-materials via Markov Chain model

For the purpose of prediction analysis, we flagged each and every user with either "0" or "1". A user who did not download any teaching-material at t-month (i.e. the observational period) was flagged by "0". Otherwise, the user received a "1"-flag. That being said, there is a blank status that represents users who never downloaded teaching-materials before the t-month.

We classify users into three categories to ascertain if there are any differences, as follows

- early users: to represent users who did the first download in between July 2021 and January 2022
- post users: to represent users who did the first download in between February 2022 and June 2022
- all users: to represent all users within the aforementioned-observational period

Here, a data point moves from t to t+1, with the unit (or dimension) of "month". A transition probability matrix is then created, given the aforementioned-movement. We give an example of interpretation on the Early User's probabilities as follows.

Table 2. Transition Probability by User Classification

User Classification	t	t+1	
		0	1
Early User	0	89.85%	10.15%
	1	74.77%	25.23%
Post User	0	81.16%	18.84%
	1	70.85%	29.15%
All User	0	85.60%	14.40%
	1	71.51%	28.49%

"0"("1") to represent the state of users who ("not") downloaded teaching-material on the time t

- Probability user who not download on t-month and remaining not download on t+1-month is 89.85% for early user, 81.16% for post user, and 85.56% for all user
- Probability user who not download on t-month and move to download on t+1-month is 10.15% for early user, 18.84% for post user, and 14.4% for all user
- Probability user who downloads on t-month and move to not download on t+1-month is 74.77% for early user, 70.85% for post user, and 71.51% for all user
- Probability user who downloads on t-month and remaining download on t+1-month is 25.23% for early user, 29.15% for post user, and 28.49% for all user

In the short term, the transition probabilities show that the users tend not to download. For obtaining the long term probability, we should find a steady state condition or equilibrium. One of the approaches is by repeated matrix multiplication. As a result, we obtained these following probabilities

$$\pi_{early\ user} = (88.05\%\ 11.01\%)$$

$$\pi_{post\_user} = (78.99\%\ 21.00\%)$$

$$\pi_{all\_user} = (83.23\%\ 16.76\%)$$

Such results suggest that the probability of users to remain in condition not downloaded or churn in the long term is 88.05%, 78.99%, and 83.23% for early user, post user, and all user, respectively.

4.2 Prediction of User Segmentation with RF value

Next, users' segments are built using the Recency and Frequency (RF) value of users based on their activities in downloading teaching-materials. The value is calculated on each user in every given month, which then is standardized by using Eq. (1) and (2). The results of the transition probability are presented in Table 4.

Table 4. User segmentation behavior. Results are based on RF-value and Markov Chain model

t	t+1						
	About to sleep	Champi-ons	Hibernat-ing	Loyal users	New users	Potential Loyalist	Promis-ing
(blank)		0.00%			98.78%	0.18%	1.04%
about to sleep	68.42%		24.90%		6.49%	0.19%	
Champions		77.73%		5.08%		17.19%	
Hibernating			88.97%		10.96%	0.07%	
Loyal users		23.53%		41.18%		35.29%	
new users	0.01%	0.00%			82.43%	0.48%	17.07%
Potential Loyalists	1.87%	0.85%	0.60%		3.06%	87.33%	6.29%
promising	18.19%				16.12%	0.41%	65.29%

For users who are in the state of "about to sleep" in time , their transition probabilities in *t+1*-month would be 24.90%, 6.49% and 0,19% for moving to hibernating, for becoming new users and for being potential loyalists, respectively. Their probability to stay at the current state in *t+1*-month is relatively high, i.e. 68.42%. It is interesting to notice that the probability of users to stay in their state at *t+1 month* is relatively high for all segments.

The equilibrium or steady state condition is achieved with the probability on each stage as follows::

$$\pi = (\ 11.59\%\ 0.10\%\ 26.40\%\quad 0.01\%\ 39.49\%\ 2.53\%\quad 19.88\%\ )$$

which in the long term period implies

- The probability of a user remains on about to sleep segment is 11.59%
- The probability of a user remains on champion segment is 0.10%
- The probability of a user remains on Hibernating Segment is 26.40%
- The probability of a user remains on Loyal Customer segment is 0.01%
- The probability of a user remains on New Customers segment is 39.49%
- The probability of a user remains on Potential Loyalists segment is 2.53%
- The probability of a user remains on promising segment is 19.88%

5. DISCUSSION

Emerging technologies have come and have been implemented in the public sectors since the last decades ago (Schwab, 2017). Especially in the Educational sector, the raising of the number of digital applications to serve various users is noticeably detected across nations, for instance in India, Pakistan, Philippines and Malaysia (Hong Kong Institute of Education, 2004). Such technology has mainly aimed for optimizing the coverage of supports that the government could give, as well as for accelerating the effectiveness of policies' and regulations' implementation from the upstream (i.e. central government) all the way to the downstream (i.e. end-users: schools, teachers, and students) (Bank, 2020). Particularly in widespread geographical location as well as the various levels of digital literacy amongst users, the adoption of digital technologies should be carefully thought through and should be well-designed, as such it could satisfy the user needs

Kurikulum Merdeka was born during the COVID-19 pandemic period and it has been accompanied by a digital tool for its implementation, named Platform Medeka Mengajar (PMM). Within the first quarter after launching, the organic adoption rate followed a steady pattern. The technical assistance that was given by the MoECRT, has helped to increase the adoption rate across user segments. As the pandemic gets more controlled, more offline-sessions were conducted, which resulted in the exponential growth of the users' adoption. Our experience shows that human touch in technological advancement is needed to create a momentum for users' adoption growth.

After its launch in February 2022, the team first needs to understand the current state of users' behavior and users' persona in the digital platform. The prediction of the recency and frequency (RF) values with Markov Chain model, revealed that 38% of users would be likely to churn (i.e. "about to sleep" or "hibernating") in the long run, if there are no intervention actions. Although it is relatively high, the percentage of "likely-churn" users does not ring any warning-alarm for the team. The team is aware that there are a number of room for improvements in this one-year-old-digital platform. During the first semester, the number of teaching-materials on the platform was limited. Increasing the quantity and variability of teaching-materials could be prioritized, as we noticed that most of "likely-churn" users could not find their teaching-materials of interest. We believe that the "likely-churn" users will be kept minimum by the teaching-materials improvement and other intervention activities.

We employed a hybrid model that involves RFM clustering method and probability-based- prediction model to estimate the users' behaviors in the long run, assuming constant conditions across all entities used in the model. Such a combination was implemented as well by Tarokh and Esmaelli Gokeh (2017) in manufacturing, Bureli (2019) in the tech-games industry, and Bekamiri (2020) in banking. Although the model works well, predictive modeling via Markov Chain model would be problematic when there are more states (in our case: more segments) and there are interactions across states (Carta, 2020). This modeling approach also does not take into account the possible confounding factors that may affect the users' behaviors to download teaching materials, e.g. mobile devices and a possibility of getting materials from other users. Nevertheless, the team has benefited from the model, to help formulate strategies to change users' behavior in the platform.

6. CONCLUSION

The adoption of intelligent approaches to optimize strategic decisions for various businesses has gained significance as a result of extensive developments in information technology and the creation of new concepts in this area. As a result, the goal of the current study was to provide a dynamic structure to predict retention of users based on their recency and frequency behavior. The application of this model can result in a viable method for formulating organizational strategies that are in line with user behavior. With this strategy, the company can concentrate on minimizing churn of users especially from promising segments.

In order to determine retention of users in the future on the PMM, the current work suggested a hybrid model that combines a Markov model with a RFM model. Additionally, by using this model to identify customer behavior stochastically, we may create a new model for setting up marketing campaigns to improve the acquisition and retention in the PMM platform.

ACKNOWLEDGEMENTS

We deeply thank to the PMM-technology team: Ruth Ayu Hapsari, Dzameer Dzulkifli, Putri Lestari for constructive comments and feedback; Bani Syahroni, Kalista Cendani, Jaya Wina for fruitful discussion during the study as well as for the support for translating the results into real-actionable plan; Bagoes Rahmat, Figarri Keisha, Muhammad Nasiruddin, Hana Rotinsulu, Bhaskoro Muthohar, and Septi Rito Tombe for the input of data-analytical work.

REFERENCES

Bank, A. D. (2020). *Innovate Indonesia: Unlocking Growth Through Technological Transformation*. Philippines: Asian Development Bank.

Bekamiri. (2020). A Stochastic Approach for Valuing Customers in the Banking Industry: A Case Study. *Industrial Engineering & Management Systems*. Vol 19, No 4, pp. 744 - 757

Burelli, P. (2019). Predicting customer lifetime value in free-to-play games, Data Analytics Applications in Gaming and Entertainment. *CRC Press*. Chapter 5, 79-107

Carta A and Conversano C. (2020). On the Use of Markov Models in Pharmacoeconomics : Pros and Cons and Implications for Policy Makers. https://www.frontiersin.org/articles/10.3389/fpubh.2020.569500/full

EsmaeiliGookeh. (2017). A New Model to Speculate CLV Based on Markov Chain Model. *Journal of Industrial Engineering and Management Studies*. Vol. 4 No.2, 2017, pp 85-102

F. Gebali. (2008). Analysis of Computer and Communication Networks. *Springer Science Business Media*. 123-124

Hughes, A.M. (1994). *Strategic Database Marketing*. Probus Publishing: Chicago, IL, USA

H. E. Roman and M. Porto. (2008). Fractional Brownian motion with stochastic variance: Modeling absolute returns in stock markets. *International Journal of Modern Physics, C, Physics and Computers*. 19, 1211-1242.

H. Roshan and M. Afsharinezhad. (2017). The New Approach in Market Segmentation by Using RFM Model. *Journal of Applied Research on Industrial Engineering*. Vol 4, No.4, 259-267

Hong Kong Institute of Education. (2004). *Reform of Teacher Education in the Asia-Pacific in the New Millennium: Trends and Challenges*. Netherlands: Springer Netherlands.

Ja'ashan, M.M.N. H. (2020). The Challenges and Prospects of Using E-learning among EFL Students in Bisha University. *Arab World English Journal*. 11 (1) 124-137

Khajvand et al. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study. *Procedia Computer Science*. 3, 57-63

Komorowski et al. (2010). On Ergodicity of Some Markov Processes. *Annals of Probability*, 1401-1443

Lumsden et al. (2008). *Customer value in an all-inclusive travel vacation club: An application of the RFM framework*. J. Hosp. Leisure Mark.

Mustakim and Syaifullah. (2015). Pengembangan Aplikasi Prediksi Penyakit Berbahaya Di Provinsi Riau Berdasarkan Model Markov Chains. *Jurnal Rekayasa Dan Manajemen Sistem Informasi*, 10-12

N. AbdulRazak et al. (2022). Prediction of Urban Spatial Changes Pattern Using Markov Chain. *Civil Engineering Journal*. Vol 8, No. 4

P. Zhi-Hang, X. Le-Tian and S. Yong-Mei. (2005). Application of Weighted Markov Chain in the Prediction of Year's Harvest of Crops. *Mathematics in Practice and Theory*, 30-35

Q. L. Song and C. Song. (2009). An Application of Markov Chain in the Prediction of the Market Economy. *Journal of Business Research,* 2, 46-49

Schwab, K. (2017). *The Fourth Industrial Revolution*. Switzerland: Penguin Books Limited.

Srikanth P. (2015). Using Markov chains to predict the natural progression of diabetic retinopathy. *Int J Ophthalmol*, 8, 132-137

Syahrul et al. (2022). The Implementation of Online Learning at The Faculty of Engineering, State University of Makassar in Response To Covid-19. *Indonesian Journal Of Educational Studies (IJES)*

Tarokh, M. J and Esmaeili, G. M. (2017), A stochastic approach for valuing customers, *International Journal of Software Engineering and its Applications*, 9(3), 59-66.

Vavoula, G. & Sharples, M. (2008) Challenges in evaluating mobile informal learning. *In Proceedings of the mLearn 2008 conference* (pp. 296-303). UK: Wolverhampton