Knowledge pricing structures on MOOC platform – A use case analysis on edX

Indicate Submission Type: Completed Research Paper

Yingnan Shi
Research School of Management
LF Crisp Building 26C
ACT, Australia, 2601
yingnan.shi@anu.edu.au

Xinghao Li
Research School of Computer Science
108 North Road
Acton ACT 2601
Xinghao.li@anu.edu.au

Armin Haller
Research School of Management
LF Crisp Building 26C
ACT, Australia, 2601
armin.haller@anu.edu.au

John Campbell
Research School of Management
LF Crisp Building 26C
ACT, Australia, 2601
john.campbell@anu.edu.au

Abstract

University courses provided online in form of MOOCs (Massive Open Online Courses) are gaining increased attention, yet their pricing structure is rarely studied. MOOCs can be treated as knowledge products, and MOOC platforms, therefore, become the marketplace for market-participants to trade those products. A functional knowledge market cannot be established without an appropriate and reliable pricing model, but so far, there have only been a very limited number of studies focusing on the pricing strategies in MOOCs. This study fills this gap by providing a systematic price analysis on one of the largest non-for-profit MOOC platforms, edx.org. In doing so, we establish a model to explain the price differences among different courses. This study can act as a well-grounded starting-point for future MOOC-pricing studies and knowledge products' valuation research.

Keywords: *Pricing Strategy; MOOCs; Knowledge Pricing; Knowledge Market; edX*

Introduction

Although online learning (also named "remote learning" or "e-learning") is not a new concept, the advent of MOOCs is still being considered by many people as a "disruptive innovation" and as a "black swan" in the higher education environment (Aparicio et al. 2014; Ryan and Williams 2014). MOOC is seen not only as a way of offering unique kinds of courses, but also as a social movement and a new attempt in the higher education sector to achieve effective distance learning through "massive usage" facilitated by the Internet, which is empowered by its collectivism, high connectivity, great openness, pronounced emphasis on interactions and pro-knowledge-sharing attitudes among peers. (Belleflamme and Jacqmin 2016; McAuley et al. 2010; Rosenberg 2005)

The first MOOC appeared in 2008, but it was not until 2012 that a small number of MOOC initiatives gained "massive" popularity worldwide (Moe 2015). Thus 2012 was therefore considered by some as "the year of the MOOC" (Pappano 2012). As Rustam and van der Weide (2016) argued, those platforms offering MOOCs, as well as the relevant software mediating and connecting the students and the course providers, can be seen as "marketplaces for knowledge" which allows knowledge buyers and sellers to

meet with each other and negotiate transactions. Ideally, this marketplace of knowledge would optimise the distribution of knowledge resources and maximise the utility level for all market participants.

A well-functioning market cannot survive without an appropriate pricing mechanism. Yet, the pricing system and the pricing strategies used by major modern MOOC platforms such as edX, Coursera, Udemy, Udacity, and xuetangX, are still under-developed. Until recently there has also been limited research investigating factors influencing the pricing of MOOCs (Jia et al. 2017). This has become an increasing concern to providers, particularly since after 2015 some of the major providers of MOOC courses have gradually dropped their free offerings (MoocLab 2017). Now, most MOOC providers charge a price to students (customers) to enroll (rather than to audit) and to obtain a trackable certificate. For instance, in 2015 edX ceased its Honor Code Certificates to its users who take courses for free (Agarwal 2015), and, in the same year, edX's major competitors, Coursera and Udacity, "phased out" most of their free "Verified Certificates" and "Statement of Achievements". As Shah (2015) reported, in some courses, users who do not pay money are not anymore able to access the whole package of course materials (e.g. taking quizzes).

However, the pricing structure for MOOCs is still immature, largely, because it is very difficult to price digital products. Economics suggests that prices should be set at a point where the marginal benefit of the product equals its marginal cost (Mankiw 2014). MOOCs typically have marginal costs close to zero, but the fixed costs are quite high. Therefore, the fixed costs need to be spread across a large number of users; this means that more popular courses should cost less per user. However, according to Shah (2015), as of December 2015, more than half of the courses sold on Coursera are priced at exactly $49 USD. The precisely same number is also the most popular price on edX, despite differing courses' subjects, offering institutions, number of enrolled students, etc. Also, the average price of courses taught in languages other than English is lower than courses in English, which might be due to the fact that many of those non-English courses are targeted at customers in developing countries. In addition, also in 2015, Udacity adopted a "flat rate" business model by which students need to pay a fixed price, USD 200/Month for gaining a Nano-degree: a 4-12 month program designed for training technical or professional skills (Morell 2015).

Maintaining and sustaining good-quality courses can be prohibitively costly for universities, because of the high fixed costs. Despite this, we can still witness the introduction of new or re-offered courses and programs on MOOC platforms (Shah 2017; Ukueberuwa 2018), even though the revenue patterns of MOOCs are still not at all clear (Romero and Ventura 2017). Although many universities are providing MOOCs for philanthropic purposes, brand building or as a pre-admission and filtering mechanism - universities offer pathway program (mostly via credit redemption) for outstanding performers - it is still worthwhile to investigate if there is a discernable underlying pricing model within MOOC marketplaces to guide courses' pricing. Before addressing this question, it is essential to understand the current pricing structure of MOOC platforms and to investigate what factors would affect course pricing. This is what this study aims to achieve. More precisely, we attempt to investigate the aforementioned question by conducting a descriptive statistic study. This study will first examine the current price structure of an online non-for-profit MOOC site, specifically, edX.org, as it is the largest not-for-profit provider in this industry (Baker and Passmore 2016). Then, we formulate a model to examine what factors can usefully explain the differences in courses' prices. Further, we are also interested in the "usefulness" of the price data: i.e. what conclusions can be made from a given price information. Intuitively, it is conjectured that price would have a significant correlation with courses' enrolment scale, exploration rate, and completion rate. It also seems plausible that higher priced courses would receive a higher ratio of certificated enrolment compared to overall enrolment. We hypothesise that: first, the higher the price, the lower the demand (in this case, the lower the number of enrolments in a course); second, that participants pay more attention to courses they paid a larger sum of money for earning certificates. To examine those hypothesis, in this research, we collected data using a Web crawler and analyse current pricing factors based on the collected data.

The remainder of the paper is structured as follows. We will first present the background literature, followed by an introduction to the research methods in Section 2. Section 3 will present the results of our statistical analysis. We conclude in Section 4 and discuss the limitations of our study and potential direction for future work.

Background Literature

Moe (2015) summarised the history of MOOC until 2015 and argues that it is hard to usefully define the term, MOOC, which has been used by mainstream media in a conflated manner and has widely been seen as a combination of technological solutionism, disruptive innovation, and a way to "mitigate our educational crisis". He suggests that MOOC refers both to an instrumental learning model and to a social movement. That said, for the purpose of this study, we use a working definition given by Kim (2014) to define and to set up a boundary for MOOC: we will not consider MOOC as a social movement but merely as a course that has the capacity to accommodate a very large number of learners, is able to be accessed via online channels regardless of a classroom analogue, and is open for assessment and for curriculum so a learner can choose to have or not to have his or her assignment graded freely and what courses to take at will. As a result, MOOC allows on-demand accreditation and a relatively flexible learning pace.

According to Rustam and van der Weide (2016), MOOC sites and MOOC-related software create an ecosystem and a knowledge market where people can trade knowledge products on. This type of knowledge marketplace is also termed "K-mall" by Skyrme (2012). Market participants (e.g. the students as the buyers and the institutions as the suppliers) bargain with each other based on their preferences in this marketplace, attempting to approach an optimised allocation of knowledge resources. In this vein, Rustam and van der Weide (2016) offer an architecture to analyse such knowledge marketplace and argue that "market information", such as price and reputation, is fundamentally important in the bargaining process and, therefore, vital for achieving a successful and smooth matching process between demand and supply. The ultimate goal of this matching is to approach the best distribution of knowledge. However, pricing MOOCs is not welcomed by all participants, especially those who are financially limited under-educated learners from developing countries. Although MOOC platforms offer tuition waivers and tuition discounts to financially challenged students, the access to these types of discounts is still limited. As a result, some previous research argued that the benefits of MOOCs are unevenly distributed, stating that the benefits of MOOCs are not flowing to where they are most needed (Christensen et al. 2013; Evans and McIntyre 2016; Oyo and Kalema 2014). Similarly, Baker and Passmore (2016)'s work expresses frustration to the fact that MOOCs are becoming "less open" and have lost one of their original promises: i.e. "incurring zero costs to students". That said, in the long run, it is not surprising that MOOC providers and agencies try to charge a price, because the development and the maintenance of MOOCs, as well as the platforms they run on, requires a lot of financial resources, which are currently still unsustainably subsidized by venture capitalists and by universities (Cusumano 2013; Gaebel 2014).

In terms of pricing strategies, Belleflamme and Jacqmin (2016) suggest, in order to continuously provision knowledge goods without excessive use of external subsidies, MOOC sites need to adapt to a "two-sided market", i.e. a market aimed at enabling interactions between at least two groups (sides) of its end-users (Rochet and Tirole 2006). For example, typical two-sided markets are eBay and AirBnB with "two groups of users", the goods and service consumers and the sellers (suppliers) (Rochet and Tirole 2003). An asymmetric pricing structure is often suitable for this type of market, i.e. a pricing strategy aimed at charging a lower price to the group that has the strongest positive cross-side effect on the other side in order to enlarge the participants' base and to start a positive feedback loop (Belleflamme and Jacqmin 2016). In the case of MOOC platforms, this theory justifies the logic that they charge relatively low fees (compared to traditional university tuitions) to students to initiate the loop whilst charge a larger fee to suppliers for revenue. In reality, according to Kolowich (2013), as of 2013, edX.org officially uses two methods to generate revenue: either by collecting the first $50,000 USD generated by a course offered by an institution or by charging the institution a course production assistance fee of $50,000 USD. In terms of business models, most current MOOC platforms are using a Freemium model (accessing a course for free while charging for certification) or a Subscription model (Gassmann et al. 2014). That said, given the diversity of the MOOC education market, there are also other possible business models, such as the Advertising model, the Subcontractor model, the sub-licensing model and the Job matching model. Those models may have a promising future according to Belleflamme and Jacqmin (2016) and Jia et al. (2017). In practice, courses, together with their certificates, are often sold in bundles (e.g. in form of a micro master program on edX). Taking that into

consideration, Jia et al. (2017) present a mathematic model for pricing the bundled certificates. As such, there does exist some sort of pricing structures on MOOC platforms, but the current system's effectiveness is at most questionable, because, according to Kolowich (2013), major MOOC players are still facing challenges in maintaining a free cash flow and a healthy financial return. For the time being, the platforms' survival is still heavily relying on external investments and cross-financing from universities. For example, Shah (2016) mentioned that, to survive, edX obtained $15 million USD from its university partners in 2015 and would gain another $20–30 million in 2016. It is, however, uncertain to see whether this model is viable in the long run.

Although many previous studies investigated the causal factors of MOOCs' low completion rate issue (Aboshady et al. 2015; Lewin 2013; Yuan et al. 2013), very few papers focus on the potential relationship between the price and the completion rate. Burd et al. (2015)'s work mentioned money and price issues, but it is only concerned with the course suppliers' perspective. Their argument is that the low completion rate affects the revenue flow generated by MOOCs, but not how a course's price would affect completion rates of students.

Recently, there has been an increase in publications for MOOC-related topics (Zhu et al. 2018), and yet, based on a Google Scholar search (keyword: pricing on MOOCs; timeframe: after 2015) we see only a very limited number of papers (even fewer high-cited papers) that discuss the pricing strategies and/or the price structures of MOOCs offered by major platforms. Also, as argued by Daradoumis et al. (2013), the number of studies using data analysis methods on existing MOOC platforms is rather limited. The reasons include that the platforms only offer limited data access and the usability to analyse the data is even lower.

Research Method

The software we developed for the project[1] consists of two major parts: a Web crawler and a data analyser. Both are written in Python 3.6, and both are specifically designed for crawling and analysing data on edX. We utilised the "selenium.webdriver" module to access the price information, as the information is not directly crawl-able: They are generated by a piece of JavaScript code embedded rather than in plain HTML format. As of 1st February 2018, the crawler visited 1743 courses' web pages and collected information about their recommended course length and suggested efforts (weekly study-load), institution, subject, level, language(s), and video transcript languages. It is worth to mention that there are 1923 courses offered on edX, but some courses are "uncrawlable" because their "courses information" section does not provide useful data fields, which may be due to the fact that the courses have ceased. Furthermore, although some courses do have a sufficient information section, the data shown in the section are not entirely structured: Most courses contain eight data fields, but some contain only seven fields while some have nine fields. Some courses have no "Price" information and some have an extra field named "type". The "type" field usually contains miscellaneous data: For instance, the course "Deals in Project Finance: Case Studies and Analysis" has a "Type" which is labelled "Professional Education and Self-Paced".

In total, we found 673 courses that currently do not contain clear price information, of which 91 courses do not have a price data field, while 582 courses are labelled as "Free Verified Certificate option closed". Other than that, we also found 42 courses being clearly labelled "free". In our following analysis, the former two categories of courses are excluded, while the third group, the "free courses", are denoted "zero" in the price column of our data-frame. In addition, there exist courses with no "language" information and courses with incomplete information for "length". In such cases, we label them as "NaN" in the dataset (i.e. not being excluded). For the purpose of data analysis, NaN, originally standing for "Not a Number", refers to a "missing data" in this study. We include those data points from our analysis as we deem their level of influence as negligible. Furthermore, as we observed that most of the "length" and "effort" data are in a format of "*n* to *m* weeks" (e.g. 2 to 5 weeks) or "*n* to *m* hours

[1] The project is called edx_crawler and is accessible https://github.com/Xenonshi/Edx_Crawler_V0.git

per week" (e.g. 10-12 hours per week), we parsed and split those data points into three columns: max length (effort), min length (effort), and average length (effort).

We also acquired a second-hand dataset which contains data for 290 courses offered by Harvard and MIT, which encompasses subjects in STEM (Science, technology, engineering, and mathematics), CS (Computer Science), GHSS (Government Health Social Sciences) and HHRDE (Humanities, History, Religion, Design, Education). We performed an inner join on those two datasets using the "URL", "course title", and "course number" as keys. We deleted those courses that are not open as of 2nd February 2018 nor provide effective price information on its course introduction webpage. Many courses have also been offered already multiple times, which is readily identifiable as those course entries that have exactly the same URL but different course launch date. For those courses, we only select the most recent one for our research. In the end, 64 courses with active price information are identified and are used for further analysis. Also, for the purpose of this study, we follow Chuang and Ho (2016)'s convention and define that a student becomes a participant after he or she enrolled a course. If a person visits more than half of the course materials, the person becomes an explorer, and the course's exploration rate can be calculated by dividing the number of explorers by the number of participants. Then, if a person earns a certificate via successfully passing quizzes, assignments and exams, the person becomes a certificated student for that course, and by dividing this number against the number of the total participants, we obtain the course' completion rate.

Results and Modelling

This section will first provide a series of basic descriptive statistic results based on our data analysis. Because all the price data crawled used Australian Dollars, unless it is otherwise noted, the denomination of price is AU Dollars (AUD). At the time of writing the paper, the exchange rate of AUD to USD is ~0.78. A basic analysis of the price of courses is shown in Table 1. It can be noted that the price frequency distribution is right-skewed with its median lower than its mean.

Measure	Price		Percentile	Price
count	1070.00		5%	31. 00
mean	115.33		25%	61. 00
std	97.28		50%	93. 00
max	1115.00		75%	123.00
min	0		95%	248.55

Table 1 Basic price statistics

In terms of "institutions vs. price", as of February 1st, 2018, there are 115 institutions offering courses, and the provider offering the highest number of courses is, somehow surprisingly Microsoft, which has so far offered 183 courses (either explicitly for free or for a price). The average price is 123.01 AUD (95.57 USD) with a small standard deviation of 0.15. HarvardX and MITx follow Microsoft, providing 68 and 47 courses with an average price of 95.29 AUD (74.03 USD) and 102.49 AUD (79.62 USD) respectively. Also, it can be observed that, while on average the lower-priced courses are not necessarily provided by an institution from a developed country, the most expensive courses are all from developed economies, especially from the US and Australia. The latter finding coincides with Ortiz et al. (2015)'s report which mentions that the US and Australia have also the most expensive (offline) university fees.

In terms of price differences for different subjects, on average, the most affordable subject is "History" while the most expensive subject is "Medicine". Details can be found in Figure 1. The online courses' fee structure seems to be very similar to fee discrimination in the offline world, as in Australia, the government implements a "Student contribution banding and ranging system" in which those "Band 3 degrees" (e.g. medicine, laws, economics, and accounting) are more expensive than "Band 2 degrees" (e.g. computing and engineering) and "Band 1 degrees" (e.g. humanities social studies, and education) (Soutar and Turner 2002; StudyAssist 2016).

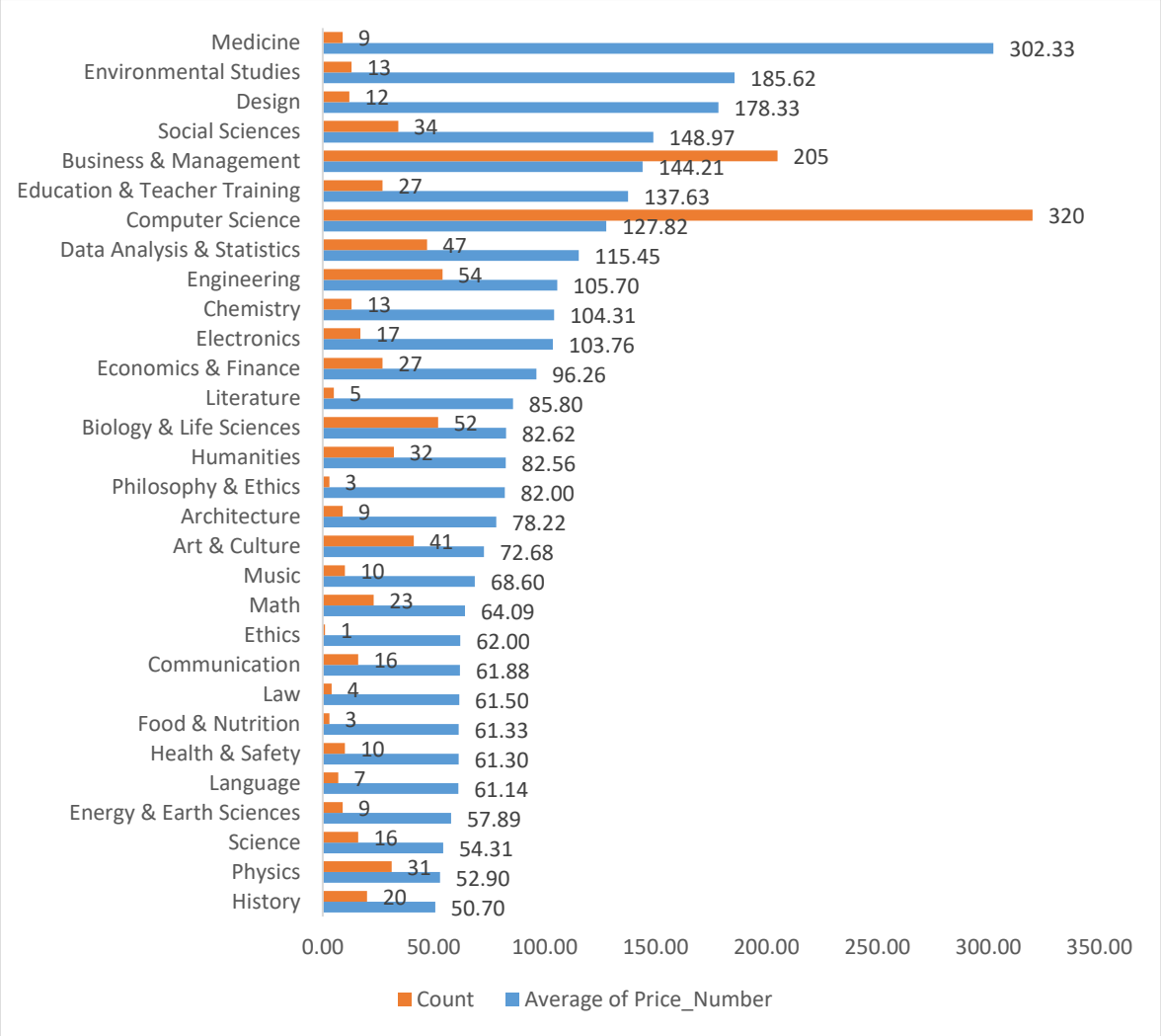


Figure 1 Price vs. Subjects

Level	*Average price*
Advanced	*208.2*
Intermediate	*123.83*
Introductory	*82.77*
Average	***115.33***

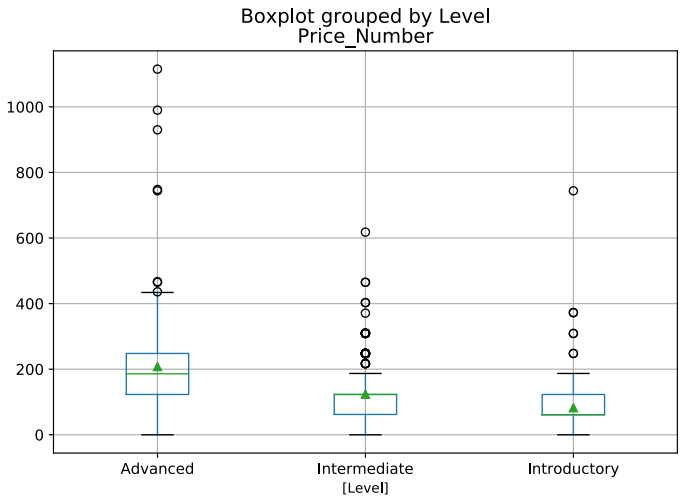


Figure 2 Average price for different levels

We also see a pattern that, on average, the more "difficult" the course, the more expensive. (See Figure 2 Average price for different levels) The average price for advanced level courses is AUD 208.2 (USD 161.42), for intermediate level courses it is AUD 123.83 (96.20 USD), and for introductory level

courses it is AUD 82.77 (USD 64.30). However, this correlation does not hold for all subjects. For example, the average price for an "Intermediate" level "Architecture" course is slightly less expensive than the "Advanced" level course in the same subject. The cost of an average "Intermediate" level course in Medicine is unexpectedly lower than the one at the "Introductory level". It can be inferred, that since the large average price difference is not only due to the course per se but also due to the offering institutions, and because the population of Medicine courses is relatively small, such price difference is exacerbated by some extreme outliers.

Language	**Price**	**Count**
English	$ 123.97	867
Japanese	$ 123.00	1
French	$ 94.55	20
English in combination with other languages	$ 92.54	74
Spanish	$ 77.26	66
Dutch	$ 61.50	2
Russian	$ 61.00	4
Japanese and Chinese	$ 61.00	2
Portuguese	$ 61.00	1
German	$ 61.00	1
Arabian	$ 61.00	1
Chinese	$ 34.00	28

Table 2 Price vs. Language

Moreover, our data (See Table 2) shows that the "language" factor seems to play a significant role in explaining the difference in courses' prices, i.e. courses taught in English are more expensive than ones taught in other languages. That said, since the population size of other languages is relatively small, a valid conclusion cannot be made at this stage.

	Introductory	**Intermediate**	**Advanced**
Introductory		-9.467*** (0.000)	-14.762*** (0.000)
Intermediate	9.467*** (0.000)		-7.908*** (0.000)
Advanced	14.762*** (0.000)	7.908*** (0.000)	

Table 3 T-test for testing inter-group difference (price vs level)

Note: p-value reported in parentheses. * Significant at 10%. ** Significant at 5%. * Significant at 1%.**

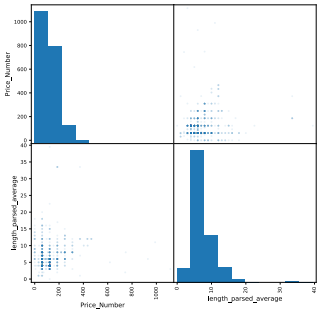
Factor	**sum_sq**	**df**	**F**	**PR(>F)**
Institution	4.328740e+06	114.0	6.264459***	0.000
Subject	1.397242e+06	29.0	5.746238***	0.000
Languages	5.009691e+05	31.0	1.738718***	0.008
Effort average	9.443550e+05	1.0	109.813258***	0.000
Length average	1.681653e+05	1.0	18.1682***	0.000

Table 4 Type II ANOVA result for factors

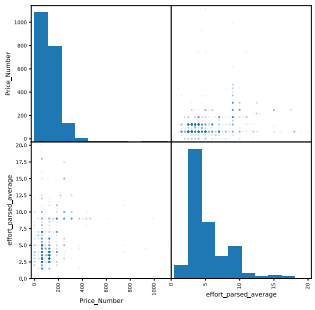
Note: * Significant at 10%. ** Significant at 5%. * Significant at 1%.**

Also, as demonstrated in Table 3 and Table 4, based on the result of an ANOVA (Type II) analysis, it is found that a significant difference exists among the prices of courses in different levels. That said, it is surprising to see that the language factor has the lowest significance level but the effort and the length have the highest. In fact, our result shows that "language" only has a significant effect at an aggregate level. If we take a closer look at the data, we found that, for each individual language, the pricing-predicating power is rather limited. For instance, the coefficient of English courses is 62.96, but the

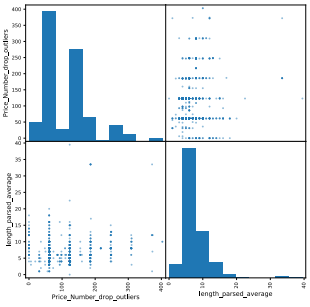
standard error is 96.463, which leads to a t-value of 0.653 and a p-value of 0.514. In contrast, the result verifies our previous assumption that some schools provide statistically more expensive courses than others. For instance, the prices of the courses provided by Curtin University and Doane University are significantly higher than the average at the 5% and 1% significance level, respectively, whilst there are universities offering significantly cheaper courses. For example, Cornell University and Edinburgh University's course price is significantly lower than the average at a 1% significance level.



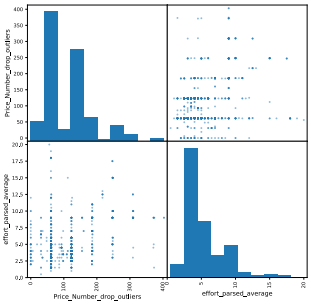
Price vs. Length



Price vs. Weekly Efforts



Price vs. Length after excluding outliers



Price vs. Weekly efforts after excluding outliers

Figure 3 Scatter plot: Price, length, and efforts

The average length and the average effort per course are 6.76 weeks and 5.16 hours per week, respectively, with a standard deviation of 3.80 and 3.04. After excluding the outliers (i.e. retaining only those that are within +/-3 standard deviations), the average length is 6.54 weeks and the average effort is 4.89 hours. Furthermore, as Figure 3 illustrates, the relationship between "price" and "course length" and "price" and "recommended weekly effort" seem to be weak, i.e. there exists no obvious relationship. Even after excluding extreme outliers in price, in length, and in weekly efforts, no clear correlation can be found.

Based on the result of our data analysis mentioned above, we propose a model to explain the relationship between the price of a course and the course's six characteristics which are used as predicting factors. We generalise this relationship to the following equation, where $Is(subject)$, $Ii(institution)$, $Ii(level)$, and $Ii(language)$ are matrixes of dummy variables. For example, $Ii(level)$ represents a course's level, and level ∈ {'introductory',' intermediate',' advanced'}.

$$price = constant + length + effort + \sum Is(subject) + \sum Ii(institution) + \sum Ile(level) + \sum Ila(language) + \varepsilon$$

We adopted the OLS (Ordinary Least Squares) method for running our regression and used a step-in strategy to formulate our modelling. It can be argued that the model gets an optimised shape at the fourth step: As Table 5 illustrated, the factor, "language", should be dropped from the model, since after

we incorporate the "language" factor into our model, the R square value has not been greatly improved, while the significant level of other factors is weakened.

	Predicator	Coef	Std.e.	|t|	P>|t|	R-square	F-value	Observations	Durbin-Watson	Jarque-Bera
Step 1	Effort	9.83	0.97	10.16***	0.01	0.11	62.63	1002.00	0.90	43916.31
	Length	2.05	0.78	2.63***	0.00					
	Constant	52.80	7.11	7.42***	0.00					
Step 2	Effort	10.40	0.94	11.12***	0.00	0.26	11.25	1002.00	1.09	34565.91
	Length	3.00	0.74	4.06***	0.00					
	Subject	-72.58	28.66	0.05	0.00					
		250.73	89.09	6.28***	0.96					
	Constant	5.97	29.01	0.21	0.84					
Step 3	Effort	12.70	1.11	11.49***	0.00	0.58	8.51	1002.00	1.45	91211.93
	Length	4.23	0.72	5.85***	0.00					
	Subject	-41.47	26.87	0.01	0.01					
		95.95	72.91	2.80	0.99					
	Institution	-231.50	0.00	0.02	0.00					
		693.37	75.00	12.72***	0.98					
	Constant	-0.51	36.17	0.01	0.99					
Step 4	Effort	9.23	1.15	8.03***	0.00	0.62	9.57	1002.00	1.49	77780.15
	Length	3.36	0.70	4.77***	0.00					
	Subject	-24.84	25.86	0.002	0.00					
		102.10	70.26	12.95***	1.00					
	Institution	-240.38	0.00	0.05	0.00					
		680.75	73.94	3.08***	0.96					
	Level[T.Intermediate]	-47.87	7.25	6.60***	0.00					
	Level[T.Introductory]	-64.02	7.62	8.40***	0.00					
	Constant	75.33	35.94	2.10**	0.04					
Step 5	Effort	9.29	1.19	7.80***	0.00	0.62	7.99	999.00	1.48	81043.44
	Length	3.63	0.73	4.95***	0.00					
	Subject	-7.74	31.40	0.06	0.02					
		127.49	47.50	2.36**	0.95					
	Institution	-241.80	0.00	0.01	0.00					
		680.70	78.80	12.80***	0.99					
	Level[T.Intermediate]	-48.27	7.41	6.52***	0.00					
	Level[T.Introductory]	-64.68	7.84	8.25***	0.00					
	Language	-99.79	0.00	0.01	0.34					
		29.92	117.55	0.95	1.00					
	Constant	89.16	85.75	1.03	0.30					

Table 5 Result of OLS Regression: Step in

Note: * Significant at 10%. ** Significant at 5%. * Significant at 1%.**

We assumed in the introduction that popular courses should cost less per user as there is a larger user base to share the fixed costs with, which may potentially create barriers to entry, for instance, for starting new courses with uncertain popularity. However, the result showed that this assumption is not valid. Table 6 shows a significantly positive relationship between the price and participants of the courses offered by Harvard and MIT on edX.org. The explanation for this unexpected result is that, according to Armstrong (2016), as in the higher education sector, education is often treated as a credence good where the "high price" implies "high quality" which attracts "high enrolment": Enrolling in a MOOC is free of any charges after all. Thinking further, higher education, even being provided online, seems to be treated by its users as Veblen Goods. As an idea proposed by Salmon (2009), for offline courses, an increase in tuition partially raises the enrolment number as the school becomes more desirable in terms of being a symbol of high status.

	Coefficient	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
I	11313.38	1293.25	8.75	0.00	8756.22	13870.53	8756.22	13870.53
Price	48.42	18.42	2.63	0.01	12.00	84.85	12.00	84.85

Table 6 Price and participants, *I* stands for intercept

Nonetheless, Table 7 and Table 8 shows that both the exploration rate and the completion rate show no significant differences when there are differences in prices. Admittedly, not all the students enrolled want to earn a certificate. In fact, according to Chuang and Ho (2016), based on a questionnaire for 195 courses offered by MITx and HarvardX, on average, only 54% of the participants self-reported an intention to obtain a certificate. Meanwhile, this divide in intentions may have a negative impact on the meaningfulness of the completion rate and exploration rate, and consequently, negatively affect our

correlation analysis. We argue that the "ratio of the number of participants who completed all course requirements against enrolled population" may be a better indicator than "certificated rate" in terms of examining completion rate. However, Harvard and MIT did not publish those relevant data in detail. Also, putting price aside, we do understand there are many other technological factors and human agency elements that might affect the exploration rate and completion rate (Chu and Robey 2008).

	Coefficient	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
I	19.51	2.66	7.33	0.00	14.19	24.84	14.19	24.84
Price	0.01	0.03	0.52	0.61	-0.04	0.06	-0.04	0.06

Table 7 Price and exploration rate, *I* stands for intercept

	Coefficient	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
I	5.86	0.94	6.25	0.00	3.99	7.74	3.99	7.74
Price	0.01	0.01	0.63	0.53	-0.01	0.02	-0.01	0.02

Table 8 Price and completion rate, *I* stands for intercept

In addition, we also notice that, for MOOCs provided by Harvard and MIT, those CS (Computer science) and STEM (Science, Technology, Engineering and Mathematics) courses tend to attract more participants than non-CS and non-STEM courses. Therefore, in order to check whether the correlation between price and enrolment is still valid when a course's subject is under consideration, we categorise the courses based on their curricular areas (either a CS&STEM course or a non-CS&STEM course). As shown in Table 9 and Table 10, it seems that the positive relationship is still valid between the price and the enrolment, though the effect of this relation is slightly smaller for those non-CS&STEM courses, as indicated by a smaller coefficient.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
I	11303.97	1299.64	8.70	0.00	8733.86	13874.08	8733.86	13874.08
Price	51.60	19.09	2.70	0.01	13.86	89.35	13.86	89.35

Table 9 Price and CS&STEM courses' participants, *I* stands for intercept

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
I	11313.38	1293.25	8.75	0.00	8756.22	13870.53	8756.22	13870.53
Price	48.42	18.42	2.63	0.01	12.00	84.85	12.00	84.85

Table 10 Price and non-CS&STEM courses' participants, *I* stands for intercept

Finally, we used a polynomial regression to examine the correlations between price and percentage of students who hold a bachelor degree or up (i.e. bachelor rate) and between price and the percentage of students who explored the course. In doing so, we first rounded the independent variable to the nearest integer value, then the average of price in each bucket to remove the spikes. In Figure 4, red dots are the actual data, blue dots are the refined data and the green curve is the output polynomial function. We chose polynomial degrees of 2, 3, and 4. These numbers are chosen as in this stage we only have a dataset with limited tuples, so choosing a higher degree may overfit the regression curve. Moreover, our result indicates that courses with either relatively very high proportion (85%) of highly educated people (bachelor or higher) or low proportion (55%) would have a lower-than-average price, whilst courses that have a bachelor rate at about 65-70% may ask for a higher price. Also, using the same method we noticed that courses with an explored rate of 20-30% have a higher price (See Figure 5). However, it requires a deeper research (e.g. using a more general and comprehensive dataset) to validate this phenomenon. At the current stage, most of the courses in our current dataset are free, impairing our training model.

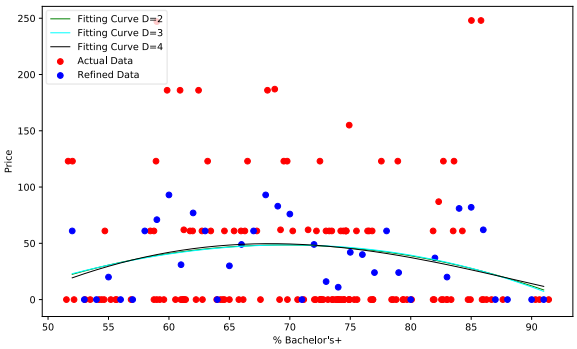


Figure 4 Price vs Percentage of students with a bachelor's and up

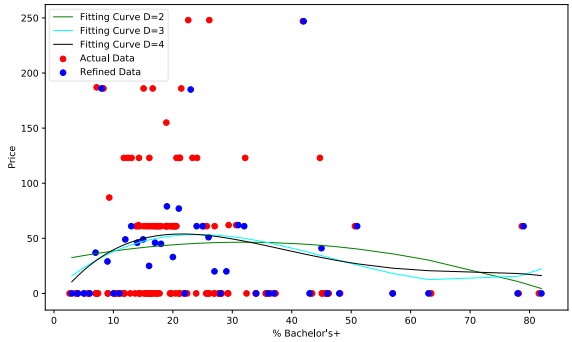


Figure 5 Price vs Explored rate

Conclusions and discussions

In order to systematically study the current price structure of the knowledge products sold on a popular knowledge market, i.e. edX.org, this study used crawlers to collect data and performed a descriptive study based on the collected data. We examined the determinant factors for course prices and formulate a model for the prices. It is found that factors, such as effort level, difficulty level, recommended length, offering institutions, and subject, are significantly correlated with courses' price but that "language" only has a significant effect at an aggregate level. We also use a fraction of courses' data to test if there exists a correlation between the prices and the courses' completion rates, but we only found that the price shows a positive correlation with participants' number but no clear relationship with exploration rate nor completion rate.

Due to the limitations in terms of available data and scope, there are several limitations that need further exploration. First, we did not address the multicollinearity problem, which requires procedures such as lasso and ridge regression to remedy or an increase of the dataset size. Second, the existing price structure established by edX, as well as its partners, may not be optimised. Instead, currently, most of the courses' prices are either pre-determined by the platform or follow a certain set of conventions rather than a price determined by the market. To maximise the welfare of all the market participants, we presume that prices for each course would be optimised if the price is set at a level where the highest Willingness-To-Pay (WTP) of a majority of the students meets the lowest Willingness-To-Accept (WTA) of the corresponding course providers. A future study shall devise methods to determine the real WTP of the consumers and the WTP of the course suppliers. Of further interest would be to find and to cluster consumer characteristics that, to some extent, determine or be able to predict the level of WTP and of WTA. Third, the course list on edX is changing, and the research conducted here is based on a snapshot of the data only. It limits the quality of and the depth of our price analysis due to the lack of time-series data. We are planning to collect this time-series data through regular future Web crawls.

We expect this research to be beneficial for MOOC platforms, online learning and e-learning industries, and relevant research areas. We acknowledge that these industries are undergoing major changes both in terms of scope and of scale. For example, in the distance learning sector, there is a notable shift in focus from developed countries to developing countries (Cheney 2017; Nataf 2018). There is also a challenge in that most providers are from the Western higher education sector, but a large potential customer base lives in developing countries where high-quality education is highly demanded, but who have only very limited access to MOOCs. Hence, it would be useful if we can design a suitable MOOC pricing structure which can better serve the learners from developing countries. We expect that this study provides a starting point for future research into discriminative pricing practices. Also, as we suggested earlier in this paper, we argue the current pricing of many MOOCs may be suboptimal. Therefore, an empirical analysis of the current MOOCs' pricing structure will support the development of a more advanced pricing mechanism that can help both the MOOC providers and platforms' owners to achieve a more accurate price-setting system and, as a result, to make the MOOC platforms become a better knowledge marketplace to allocate the knowledge resources in a more-optimized manner. Henceforth, the idea to correlate price and statistics data is useful to design an algorithm to dynamically adjust the price of the future courses based on the previous statistics to maximize the gross revenue, a potential direction for future work.

References

Aboshady, O. A., Radwan, A. E., Eltaweel, A. R., Azzam, A., Aboelnaga, A. A., Hashem, H. A., Darwish, S. Y., Salah, R., Kotb, O. N., and Afifi, A. M. 2015. "Perception and Use of Massive Open Online Courses among Medical Students in a Developing Country: Multicentre Cross-Sectional Study," *BMJ open* (5:1), p. e006804.

Agarwal, A. 2015. "News About Edx Certificates," in: *edX learner news*. America: edX.

Aparicio, M., Bacao, F., and Oliveira, T. 2014. "Mooc's Business Models: Turning Black Swans into Gray Swans," *Proceedings of the International Conference on Information Systems and Design of Communication*: ACM, pp. 45-49.

Armstrong, L. 2016. "Barriers to Innovation and Change in Higher Education," *TIAA-CREF Institute*).

Baker, R. M., and Passmore, D. L. 2016. "Value and Pricing of Moocs," *Education Sciences* (6:2), p. 14.

Belleflamme, P., and Jacqmin, J. 2016. "An Economic Appraisal of Mooc Platforms: Business Models and Impacts on Higher Education," *CESifo Economic Studies* (62:1), pp. 148-169.

Burd, E. L., Smith, S. P., and Reisman, S. 2015. "Exploring Business Models for Moocs in Higher Education," *Innovative Higher Education* (40:1), pp. 37-49.

Cheney, C. 2017. "The Road to Real Results for Online Learning in Developing Countries." Retrieved 10th Feb., 2018, from https://www.devex.com/news/the-road-to-real-results-for-online-learning-in-developing-countries-89884

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., and Emanuel, E. J. 2013. "The Mooc Phenomenon: Who Takes Massive Open Online Courses and Why?,").

Chu, T.-H., and Robey, D. 2008. "Explaining Changes in Learning and Work Practice Following the Adoption of Online Learning: A Human Agency Perspective," *European Journal of Information Systems* (17:1), pp. 79-98.

Chuang, I., and Ho, A. D. 2016. "Harvardx and Mitx: Four Years of Open Online Courses--Fall 2012-Summer 2016,").

Cusumano, M. A. 2013. "Are the Costs Of'free'too High in Online Education?," *Communications of the ACM* (56:4), pp. 26-28.

Daradoumis, T., Bassi, R., Xhafa, F., and Caballé, S. 2013. "A Review on Massive E-Learning (Mooc) Design, Delivery and Assessment," *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2013 Eighth International Conference* IEEE, pp. 208-213.

Evans, S., and McIntyre, K. 2016. "Moocs in the Humanities: Can They Reach Underprivileged Students?," *Convergence* (22:3), pp. 313-323.

Gaebel, M. 2014. "Moocs: Massive Open Online Courses. An Update of Eua's First Paper (January 2013). Eua Occasional Papers," European University Association, Brussels.

Gassmann, O., Frankenberger, K., and Csik, M. 2014. *The Business Model Navigator: 55 Models That Will Revolutionise Your Business*. Pearson UK.

Jia, Y., Song, Z., Bai, X., and Xu, W. 2017. "Towards Economic Models for Mooc Pricing Strategy Design," *International Conference on Database Systems for Advanced Applications*: Springer, pp. 387-398.

Kim, P. 2014. *Massive Open Online Courses: The Mooc Revolution*. Routledge.

Kolowich, S. 2013. *How Edx Plans to Earn, and Share, Revenue from Its Free Online Courses*.

Lewin, T. 2013. "Master's Degree Is New Frontier of Study Online," *The New York Times*).

Mankiw, N. G. 2014. *Principles of Macroeconomics*. Cengage Learning.

McAuley, A., Stewart, B., Siemens, G., and Cormier, D. 2010. "The Mooc Model for Digital Practice,").

Moe, R. 2015. "The Brief & Expansive History (and Future) of the Mooc: Why Two Divergent Models Share the Same Name," *Current issues in emerging elearning* (2:1), p. 2.

MoocLab. 2017. "State of the Mooc 2016: A Year of Massive Landscape Change for Massive Open Online Courses," Online Course Report.

Morell, C. 2015. "Udacity Nanodegree Reviews: Your Questions Answered," Udacity.

Nataf, E. 2018. "Education Technology Is a Global Opportunity," Crunch network.

Ortiz, A., Chang, L., and Fang, Y. 2015. "International Student Mobility Trends 2015: An Economic Perspective," *World Education News & Reviews*).

Oyo, B., and Kalema, B. M. 2014. "Massive Open Online Courses for Africa by Africa," *The International Review of Research in Open and Distributed Learning* (15:6).

Pappano, L. 2012. "The Year of the Mooc," *The New York Times* (2:12), p. 2012.

Rochet, J. C., and Tirole, J. 2003. "Platform Competition in Two-Sided Markets," *Journal of the european economic association* (1:4), pp. 990-1029.

Rochet, J. C., and Tirole, J. 2006. "Two-Sided Markets: A Progress Report," *The RAND journal of economics* (37:3), pp. 645-667.

Romero, C., and Ventura, S. 2017. "Educational Data Science in Massive Open Online Courses," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (7:1).

Rosenberg, M. J. 2005. *Beyond E-Learning: Approaches and Technologies to Enhance Organizational Knowledge, Learning, and Performance*. John Wiley & Sons.

Rustam, D. D., and van der Weide, T. P. T. 2016. "The Knowledge Market Online Learning (K-Mall) Architecture for Higher Education," *Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on*: IEEE, pp. 135-140.

Ryan, P., and Williams, A. 2014. "The Commercialisation of Moocs," *12th APacCHRIE Conference*, pp. 21-24.

Salmon, F. 2009. "It's Expensive, So It Must Be Good," in: *The Economist*.

Shah, D. 2015. "Mooc Trends in 2015: The Death of Free Certificates," Class Central.

Shah, D. 2016. "Edx's 2016: Year in Review."

Shah, D. 2017. "200 Universities Just Launched 600 Free Online Courses.," Quartz.

Skyrme, D. 2012. *Capitalizing on Knowledge*. Routledge.

Soutar, G. N., and Turner, J. P. 2002. "Students' Preferences for University: A Conjoint Analysis," *International Journal of Educational Management* (16:1), pp. 40-45.

StudyAssist. 2016. "Australian Government Study Assiststudent: Student Contribution Amounts ", Study Assist, Australian Government Study Assist.

Ukueberuwa, M. 2018. "Return of the Moocs: An Innovative Nonprofit Shows How Internet-Based Education Can Help Contain the Spiraling Costs of College.," reManhattan Institute for Policy Research.

Yuan, L., Powell, S., and CETIS, J. 2013. "Moocs and Open Education: Implications for Higher Education."

Zhu, M., Sari, A., and Lee, M. M. 2018. "A Systematic Review of Research Methods and Topics of the Empirical Mooc Literature (2014–2016)," *The Internet and Higher Education*).