

Towards triggering higher-order thinking behaviors in MOOCs

Xu Wang

School of Computer Science
Carnegie Mellon University
5000 Forbes, Pittsburgh, PA, 15213
xuwang@cs.cmu.edu

Miaomiao Wen

School of Computer Science
Carnegie Mellon University
5000 Forbes, Pittsburgh, PA, 15213
mw@cs.cmu.edu

Carolyn P. Rosé

School of Computer Science
Carnegie Mellon University
5000 Forbes, Pittsburgh, PA, 15213
cprose@cs.cmu.edu

ABSTRACT

With the aim of better scaffolding discussion to improve learning in a MOOC context, this work investigates what kinds of discussion behaviors contribute to learning. We explored whether engaging in higher-order thinking behaviors results in more learning than paying general or focused attention to course materials. In order to evaluate whether to attribute the effect to engagement in the associated behaviors versus persistent characteristics of the students, we adopted two approaches. First, we used propensity score matching to pair students who exhibit a similar level of involvement in other course activities. Second, we explored individual variation in engagement in higher-order thinking behaviors across weeks. The results of both analyses support the attribution of the effect to the behavioral interpretation. A further analysis using LDA applied to course materials suggests that more social oriented topics triggered richer discussion than more biopsychology oriented topics.

Author Keywords

Discussion; Learning analytics; LDA topic modeling; regression analysis; coding manual; propensity score matching;

ACM Classification Keywords

K.3.1 Computer Uses in Education

1. INTRODUCTION

Previous work in the field of CSCL (Computer Supported Collaborative Learning) has demonstrated that discussion can facilitate learning in contexts such as classrooms or intelligent tutoring systems [7,9]. However, unlike traditional educational settings, discussions in MOOCs (Massive Open Online Courses) are large-scaled and asynchronous in nature, and therefore more difficult to control. In order to design interventions to increase the quality of discussion in MOOCs, we begin by investigating the following two research questions. 1) What kinds of discussion behaviors are associated with more learning? 2) What kinds of learning materials appear to trigger more of these discussion behaviors? These findings have the ability to inform development

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

LAK '16, April 25-29, 2016, Edinburgh, United Kingdom
© 2016 ACM. ISBN 978-1-4503-4190-5/16/04...\$15.00
DOI: <http://dx.doi.org/10.1145/2883851.2883964>.

of interventions to improve discussion and learning in MOOCs.

Driven by the first research question, we conducted an extension of a previous study to explore the relationship between cognitively relevant discussion behaviors and learning. In [26], we developed a coding scheme based on Chi's [8] ICAP (Interactive-Constructive-Active-Passive) framework to categorize students' cognitive engagement displayed in their conversations. In the exploratory study, we found that on-topic discussion is correlated with more learning than off-task discussion. However, contrary to Chi's earlier findings, we did not find a demonstrable order of strength between interactive, constructive and active discussion behaviors as we had expected.

One potential explanation is that the previous version of the coding manual was a low inference coding manual that may have missed subtle nuances that are germane to the distinctions made in the original ICAP work. For example, one case that posed a challenge for the low inference approach was distinguishing cases where students describe a life event for the purpose of explaining a psychological concept from cases where they are just socialized, or cases where they are explaining to someone else vs. explaining to themselves. In this paper, we present a higher inference approach to coding designed to overcome these difficulties.

Despite the increased difficulty of achieving high reliability with a high inference coding manual, we achieved a high reliability after a 3 month concerted effort. Two coders coded 40 randomly sampled posts, and achieved Kappa of 0.721 and weighted Kappa of 0.864. The coding manual makes fine-grained distinctions between types of active, constructive, and interactive behavior. However, constructive and interactive behaviors are rare. Therefore, in our analysis building on this coding, we aggregate constructive and interactive behaviors together, referring to them as higher-order thinking behaviors, which involve constructing ideas that go beyond what is explicitly covered in the course materials, reasoning about course materials, and referring to peers' ideas constructively. We define active behaviors as attending to course materials in any way, including indications that the student directly repeats or paraphrases course content, or indirectly implies he/she is paying attention to course materials. We offer this coding manual as one contribution of this work to aid the work of this community in investigating research questions related to the connection between discussion and learning.

In this paper, we present an analysis in which we observe a larger effect of higher-order thinking behaviors on learning compared with attending to course materials in a less engaged manner. We then asked the question of whether the students who displayed more higher-order thinking behaviors were simply different kinds of learners who were more competent, or whether a more

stringent analysis would still suggest that these students learnt more through cognitively engaging with course materials as shown by their discussion.

To answer this question, we adopted two approaches. First, we used propensity score matching to pair students who have a similar level of involvement in other activities in the course, and compared learning between the matched control and treatment groups. Second, we performed a week-level within-subject analysis. We investigated whether individual learners had higher course performance during weeks in which they displayed more higher-order thinking behaviors in the discussion forum. Both of these analyses show that students learnt more through engaging with higher-order thinking behaviors.

Building on these results, we then investigated what is different in the course materials across weeks that lead to different levels of cognitive engagement, and what topics may trigger cognitively richer discussion among students. To answer this question, we used LDA (Latent Dirichlet Allocation) [4] topic modeling to fit a topic model in order to identify which topics are associated with the occurrence of higher-order thinking behaviors. We then follow up on this analysis with a qualitative analysis to contrast the composition of texts that are associated with cognitively richer discussion and cognitively poorer discussion.

In the remainder of the paper, we begin by briefly introducing the ICAP framework that grounds our coding manual. We also discuss current learning analytics approaches applied in a MOOC context, and studies of what kinds of materials trigger more higher-order thinking behaviors in general learning settings. In the data preparation section, we introduce the course that provided the data for this research, explain the coding manual, and introduce how we prepared our data. Next we present our analyses and results. We conclude by discussing the limitations and implications of this work.

2. RELATED WORK

2.1 Introduction to ICAP Framework

In our analytic work, we start with a foundation in the ICAP framework [8], in which the authors focus on the amount of cognitive engagement that can be detected by observation of fine-grained behaviors. The ICAP framework proposes that overt behaviors, such as discussion behaviors, can reveal students' level of engagement with learning materials. The authors explain, "Although far from perfect, overt behaviors are a good proxy to reflect different modes of engagement that teachers can use to ascertain whether a student is in fact engaged in a specific mode for a given activity." [8]

In a traditional education context, researchers have the opportunity to directly observe learners in person, which allows them to capture a larger variety of overt activities, such as students' body gestures or their nuanced interactions with their collocated peers. In a MOOC context, although we have access to a large volume of clickstream data, we lack an up-close view of student behavior illustrating their learning processes. Adopting a content analysis approach applied to student posts is one way we might estimate students' cognitive engagement in the course.

In brief, Chi and Wylie [8] propose that there are different modes or categories of "active learning", corresponding to different overt behaviors that elicit differential knowledge acquisition or learning processes. They propose that learning activities and their resulting overt engagement behaviors can be differentiated into one of four modes: interactive, constructive, active or passive.

In our work we borrow the definitions (italicized) given in the Chi and Wylie [8] paper to explain the differences between these four categories.

Passive: *Passive mode of engagement is defined as learners being oriented toward and receiving information from the instructional materials without overtly doing anything else related to learning.*

Active: *Learners' engagement with instructional materials can be operationalized as active if some form of overt motoric action or physical manipulation is undertaken.*

Constructive: *Constructive behaviors are defined as those in which learners generate or produce additional externalized outputs or products beyond what was provided in the learning materials.*

Interactive: *Interactive behaviors should meet two criteria (a) both partners' utterances must be primarily constructive, and (b) a sufficient degree of turn taking must occur.*

This framework has been widely adapted in capturing students' behaviors and conversations in learning science studies. In this work, we developed a coding manual to operationalize these behaviors in a MOOC discussion forum to better capture students' cognitive processes in these settings.

2.2 Learning analytics approaches in MOOCs

The data trace that is produced when students participate in a MOOC is multi-faceted, and specialized analytics have been developed for each type of data.

2.2.1 Clickstream data analytics

Clickstream data is produced whenever a student clicks on anything on the MOOC platform. This is the most frequent data from MOOCs that is analyzed in prior work. For example, Ferguson [12] observed seven distinct patterns of engagement in 4 MOOCs based on patterns identified in clickstream data. In [17], the authors developed a classification method to identify 4 prototypical trajectories of engagement based on students' click data. Anderson [2] found that students who participated in other platform activities (videos, quizzes, etc.) participated more in the forum as well. In [18], the authors used a machine learning method to classify students as "successful" or "risk" students based on their performance in the discussion forums, including number of posts, comments, votes, etc. These analyses provide insights into how students' general behaviors in a MOOC look like.

In [19] and [20], Koedinger et al. claimed a causal effect of learning-by-doing activities on learning by analyzing clickstream data of the psychology MOOC, the same course we did our analysis on. In this course, the clickstream data were far richer than usual because of learning-by-doing activities integrated into the platform, each of which produced their own clickstream data. The findings from clickstream data collected from the Psychology MOOC were very effective in explaining variation in student learning patterns. In our work, in order to probe more deeply into student ideas and reasoning, we explore in the same MOOC what is revealed through student discussion behavior.

2.2.2 Discourse analytics

Compared with log data, textual data is less structured, but it also contains large amount of information about student's engagement with the course. With the advances in machine learning and language technologies, development of analytic techniques to be applied to student discussion data with the aim of providing adaptive support is an up-and-coming area of learning analytics.

Allen et al. [1] used two NLP (Natural Language Processing) tools to investigate the potential of NLP techniques to perform stealth assessments of students' reading comprehension skills. Hsiao [16] proposed a novel Topic Facet Model, which treats all words in a single sentence as emanating from one topic facet. They also prototyped a visual analytics interface to present online discussion forum semantics, which helps users examine post content by viewing the topic facets. In [23], the authors used LDA topic modeling to extract topics from discussion forums, and used text mining techniques to find what are the place names students talked about most in the class, which provided insight to inform future development of the course. In [5], the authors proposed a generative model to cluster threads and rank them based on relevance to address the problem that there are far more threads than one can read.

These methods and techniques help visualize discussions in MOOCs. While there is great potential for applying machine learning methods and language technologies to improve learning environments, one downside of these automated methods is that they lose valuable information in the modeling process. In [11], in order to compensate for this problem, the authors adopted a mixed method approach, by comparing the result from machine modeling and human annotation. They did a clustering analysis of MOOC discussion posts, and also qualitatively evaluated the clusters and compared the clustering result with manual annotations, for comparison and contrast.

Several studies on discussion forums in MOOCs have examined how social factors affect attrition [24]. While study of attrition is important, one reason why so much of the recent research has focused on this issue is because of convenience. Frequently it is not possible to evaluate learning because no formal pre and post-test are administered to students. In our work we are able to explore the connection between discussion behavior and learning in a MOOC context because the MOOC we are studying included a formal pre and post-test.

2.3 Higher-order thinking behaviors

In the second half of this paper, we investigate what kinds of materials appear to trigger richer discussion. In line with this purpose, we also examine existing literature on the relationship between learning materials and discussion behaviors or cognitive engagement, especially in an online learning context.

McKendree et al. [21] suggested that the social, participatory and shared verbal activity in online environments is a trigger for higher-order thinking, because learners can see their peers and tutors modeling the process of interpretation and application; they can analyze and compare their own understanding to that of others. This suggests that MOOCs have a natural deficiency in triggering higher order thinking behaviors, because there are few opportunities for students to get intensively involved in social, participatory or shared verbal activity compared with other learning environments. This study also suggests that more work should be done to support students' participatory and shared verbal activities in MOOCs.

Zhu [27] listed many variables that may influence interaction and cognitive engagement in an online discussion, including the instructor's presence, role, discussion questions, etc. In addition, student intrinsic motivation and prior knowledge of and interest in the topic may also influence levels of cognitive engagement and interaction with peers during the discussion. Students with a higher level of prior knowledge of the subject may feel bored, while students with limited prior knowledge may find interesting in contributing to the discussion.

In addition to course-level variables, other studies show the effect of text or topics on learner's engagement. For example, McNamara et al. [22] found that readers who know little about the domain of the text benefit from a coherent text, whereas high-knowledge readers benefit from a minimally coherent text. They argue that the poorly written text forces the knowledgeable readers to engage in compensatory processing to infer unstated relations in the text.

These studies inform us about variables that could have an influence on the level of cognitive engagement triggered by learning materials, including students' interest, prior knowledge, and coherence of the text. This prior work raises questions about what we might see specifically in the differential observed engagement in cognitive behavior associated with different parts of the material in this course.

3. DATA PREPARATION

3.1 Introduction to Psychology as a science

The dataset we used in this analysis is from the course "Introduction to Psychology as a Science" offered through Coursera collaboratively by Georgia Institute of Technology and Carnegie Mellon University. The course incorporated elements from the OLI (Open Learning Initiative) "Introduction to Psychology" learning environment. One special characteristic of the course was that it administered a pre/post test with the intention to support research.

Course materials included video lectures, assigned MOOC activities, learning-by-doing activities in the OLI environment, and weekly high-stakes quizzes. We present the topics of the course by week in Table 1 to give readers a sense of what psychology topics are covered in the course.

Table 1. Topics in each week of the course

Week	Topic	Week	Topic
1	Experiment methods	2	Biopsychology, neuroscience
3	Sensation and perception	4	Learning
5	Memory	6	Language and intelligence
7	Development	8	Motivation and emotion
9	Personality	10	Social environment
11	Disorders	12	Treatment

3.2 Course involvement data

Altogether, the dataset contains data from 27,750 registered users, and a total of 7,990 posts and comments. Of these, 1079 learners have both pretest and posttest on record, 491 of whom participated in the discussion forum, generating 3864 posts and comments in total.

In addition to forum records, students' clicks related to course materials are also recorded in the clickstream data. The course record contains 1487665 student clicks. This data provides us with the opportunity to monitor students' interaction with course materials over their participation trajectory.

3.3 Coding Manual

One contribution of our work is a coding manual to capture students' discussion behaviors that are associated with learning based on Michelene Chi's ICAP framework [8]¹.

However, to operationalize discussion behaviors in a MOOC context, we need to consider specific characteristics of MOOC forums. In this coding manual, we explained how MOOC forum data is different from classroom conversation data, and how we adapted the definitions from Chi's original ICAP framework.

In a MOOC discussion forum, as long as the student contributes an on-topic post, which in our coding manual requires that students show evidence of at least attending to learning materials in some way, we consider it to be active behavior. Our definition precludes posts being identified as on-topic and passive. Thus, we do not include the passive category in our coding manual. Our definition of interactive behavior is also slightly different from Chi's framework. As students rarely take turns in conversations in the discussion forum, we hardly see "perfect" interactive activities as defined. Thus, we categorize a post as interactive if the post is constructive and the student is referring to someone else's idea expressed earlier in the conversation. Although in a MOOC context, posts often receive a nominal reply, there are nevertheless very few posts that can be considered as genuinely "interactive" in the sense intended in ICAP. Most frequently, students post to the discussion forums as a form of personal reflection, or self-explanation, without pointing to or making connection with their peers' expressed ideas. We only code interactive behaviors as such when they are constructive and show that students' are pointing to, building upon, making connections to, or challenging someone else's ideas. Thus, this code is rarely used in our coding.

As described in the introduction section, we have developed a high-inference coding manual to accurately capture student's discussion behaviors corresponding to their underlying cognitive processes. On the one hand, this means we provide definitions of each category that are content-focused instead of linguistics-focused. On the other hand, in this new version of coding manual, we provide coders with a context, including the thread starter and previous post, to better infer whether the content is on-topic, or whether the student is referring to someone else's ideas. In the previous coding manual, in which only one single post was provided for coding without further reference to its context, it can be difficult to make decisions that would depend upon knowing what the author of a post was referring to, even in identifying whether the post is on task.

A decision tree used in the coding manual is shown in Figure 1. In the first stage of the coding process, the coder is asked to decide whether a post is on-task or off-task. In the second stage, for on-task discourse, the coder is then asked to decide to which specific categories of cognitive engagement each post belongs.

Note that interactive and constructive behavior have to be course content related (in this case, psychology) in our coding. For example, a student reasoning about a scoring rubric would be categorized as displaying active behavior only, because the student is paying attention to course materials, but is not engaged in constructive behavior for the purpose of learning.

In this decision tree presented here, we provide only a brief definition of each category. In practice, we provide our coders with more detailed definitions, examples, and failed examples,

which can be viewed in the coding manual found at the URL mentioned above.

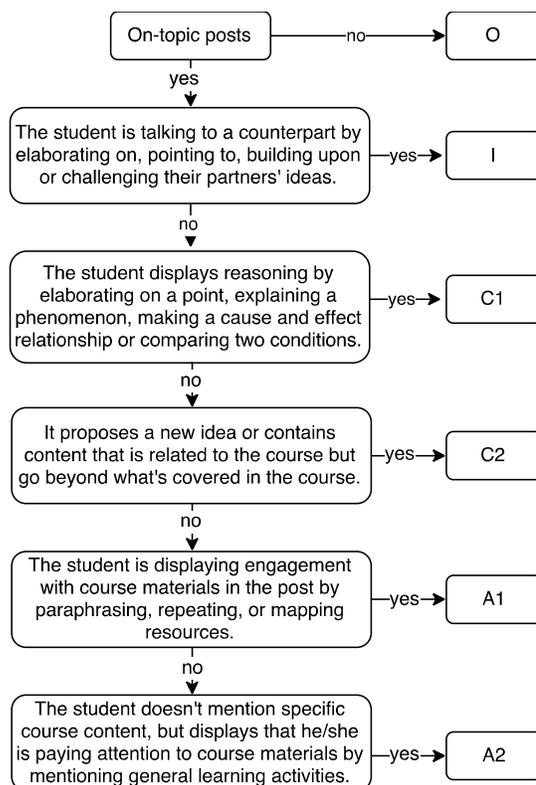


Figure 1. Decision Tree of the ICAP coding manual

3.4 Hand-coded dataset

Because a reliability test indicated high inter-rater reliability with the new coding scheme, for our analysis presented in this paper, we had one coder code all 3864 posts for students who have pretest and posttest on record in the psychology course. There are six mutually exclusive categories, O, A2, A1, C2, C1, I, as shown in Figure 1. In our analysis, we grouped A2, A1, and C2, C1 into A and C respectively.

As explained earlier, we see little intensive interaction between students in the MOOC context. Consistent with that, interactive behavior in our dataset is very rare. Thus we do not distinguish between interactive and constructive behaviors in our further analysis. Instead, we group them together, collectively referring to them as higher-order thinking behaviors. We also refer to active behavior as paying general or focused attention to course materials, as indicated in the coding manual.

4. WILL ENGAGING IN HIGHER-ORDER THINKING BEHAVIORS ASSOCIATED WITH MORE LEARNING?

In a previous study [26], we explored the relationship between cognitively relevant discussion behaviors and learning. We observed an effect of on-topic discussion over off-topic discussion, but we did not see a rank ordering in the effect of interactive, constructive, and active behaviors on learning as indicated by Chi [8]. We partly attributed that to the deficiency in the coding manual. As illustrated earlier, we revised the coding manual to more accurately capture students' discussion behaviors. In this analysis, we hypothesize that higher-order thinking

¹ <http://dance.cs.cmu.edu/MOOC-ICAP-Manual.pdf>

behaviors would have a larger effect size on learning compared with paying general or focused attention to course materials.

Driven by this hypothesis, we did a regression analysis as described in section 4.1, and observed that students who display higher-order thinking behaviors have more learning gains than those who did not display any higher-order thinking behaviors but show that they are paying active attention to course materials from their posts in the forums; And the students whose discussion behaviors indicate they are paying active attention to course materials also have higher learning gains than students who are constantly being off-topic in the forums.

We posit two possible explanations for the observed phenomenon. First, it might be that the students who displayed higher-order thinking behaviors are a different kind of learners who are better at argumentation and expressing themselves. It might be not that engaging in higher level thinking in discussion forums is causing learning gains for these learners, but rather, that such learners tend to engage at higher cognitive levels that correlates with higher learning gains. Alternately, it might be the case that by participating in higher-order thinking in the discussion forums, these students engaged with learning materials deeper, and learned more during the course as a result. While the definitive answer can only be gained through a manipulation study, we can gain stronger suggestive evidence through more fine-grained analysis. First, we adopted the propensity score matching approach to account for individual differences. We used propensity score matching to pair students who are categorized as similar learners by their involvement in other course activities, and we will describe this analysis in section 4.2. Second, we conducted a within-subject analysis over weeks to observe the effect of higher-order thinking behaviors on individual student's learning outcomes. We will describe this analysis in section 4.3

4.1 Are higher-order thinking behaviors associated with more learning: a regression analysis

The purpose of our first analysis is to measure an effect of higher-order thinking behaviors over paying general or focused attention to course materials. To this end, we created three mutually exclusive groups of students based on the highest level of cognitively relevant discussion behaviors they displayed in the forum. The three binary group variables are defined below.

Group2[higher-order]: A binary variable, which equals 1 if the student has contributed at least one constructive or interactive post during the course, otherwise it equals 0.

Group1[paying-attention]: A binary variable, which equals 1 if the student has contributed at least one active post during the course but has not displayed any constructive or interactive posts, and otherwise equals 0.

Group0[off-topic]: The group associated with this binary variable contains the rest of the students, i.e., students who have not contributed any on-topic discussion during the course.

We arranged our dataset such that the student is the unit of analysis. By including only the students who have both pretest and posttest on record, and have participated in the discussion forum, we arrive at a final sample size of 491. Below, we explain our dependent and control variables.

Dependent variable:

Post-test: The student's standardized post-test score.

Control variables:

Pre-test: The student's standardized pre-test score.

Numpost: The total number of posts the student has contributed throughout the course.

OLI-registration: A binary variable indicating whether the student has registered for OLI (Open Learning Initiative), which offered them supplementary learning-by-doing activities in each unit. [19,20]

Video: The student's number of videos clicked on at least once.

Quiz: The student's number of quizzes clicked on at least once.

OLIsite: The student's number of clicks on the OLI website.

Forum: The student's number of clicks on the forum.

We standardized all four variables based on clickstream data.

Table 2. Regression model of discussion behaviors on learning

Control/Indep. Variable	Model 1 (N=491)
Group2[higher-order]	0.456***
Group1[paying-attention]	0.264*
Pretest	0.236***
Numpost	0.049
Video	-0.015
Quiz	-0.035
OLIsite	0.043
Forum	0.019
OLI_registration	0.290*

(p<0.001***, p<0.01**, p<0.05*)

The result shows that controlling for the number of other activities the student engaged in during the course, including watching videos, doing quizzes, visiting the forum, and visiting the OLI website, for students who contributed the same number of posts, those who displayed higher-order thinking behaviors had higher learning gains than students who did not display higher-order thinking behaviors (p-value= 0.001). Similarly, students whose discussion behaviors indicated paying active attention to course materials had more learning gains than student who did not show active attention to course materials (p-value=0.031). We further contrasted between the higher-order thinking group and the paying attention group by computing the relative effect size associated with these binary variables. We found that the higher-order thinking group had an effect size of 0.36 in comparison with the off-topic group, while the paying-attention group had an effect size of 0.26. In both cases, we computed effect size using the Cohen's d method. Due to this difference in their relative effect sizes, we consider that displaying higher-order thinking behaviors in discussion forums is associated with more learning gains than displaying discussion behaviors that show general or focused attention to course materials, which in turn is associated with more learning gains than posting in the forum but being off-topic all the time. Using the new high inference coding, we see a rank of the effect of discussion behaviors in learning, as proposed in Chi's [8] framework.

4.2 Accounting for individual difference using propensity score matching

Propensity-score matching is a type of nonrandomized study that can be used to minimize selection bias and estimate the effects of treatments on outcomes [14]. It works by matching students inside the treatment group with “doppelgangers” in a comparator group. Only corresponding students with a high degree of similarity on relevant variables should be paired [3, 25]. The propensity score matching method has been recently applied to MOOCs in part because the availability of a large student population makes it feasible to identify well-matched comparator subjects [6]. In our work we use propensity score matching in order to minimize potential selection bias.

Using a propensity score matching method, we sought to match pairs of students who engaged in the same number of other activities in the course, but varied in either displaying or not displaying higher-order thinking behaviors. The students who displayed higher-order thinking behaviors are considered to be in the treatment group, while the students who did not display such higher-order thinking are in the control group. In this way we can evaluate the association between the discussion behaviors and learning, while holding other important effort related variables constant.

We used all the control variables as introduced earlier as features of the students’ engagement in the course, and built a logistic regression model to predict the students’ probability of displaying higher-order thinking behaviors in the course. We trained the model and then did propensity score matching with this model on a sample of 1079 students, which is the total number of students in the dataset that have a pre-test and post-test on record.

Table 3. Comparison between treated and control groups on relevant variables

Variable	Mean	
	Treated	Control
Numpost	7.6471	7.6555
Pretest	11.269	11.857
OLI_registration	0.85714	0.88235
Video	6.2689	6.9664
Quiz	2.9328	3.1933
OLIsite	17.504	19.857
Forum	25.58	31.697

Table 4. Regression model of treatment [higher-order thinking behaviors] on learning

Control/Indep. Variable	Model 2 (N=119*2)
Treatment	0.300*
Pretest	0.183**
Numpost	-0.091
OLI_registration	0.533**
Video	0.079
Quiz	-0.030
Forum	0.125

($p < 0.001$ ***, $p < 0.01$ ** , $p < 0.05$ *)

119 pairs of students were matched in the process. As displayed in Table 3, we compared the difference between the matched treatment and control groups. None of the features we used for the

matching are significantly different between the treatment and control groups in the matched set, which demonstrates a successful match for further analysis.

We then fitted a regression model using students’ standardized posttest score as a dependent variable, the binary treatment variable as a main effect, and the other variables as covariates, which are measurements of student’s engagement in course activities. The result is displayed in Table 4. Being in the treatment group, which indicates the student has displayed higher-order thinking behaviors in the course, has a significant effect on learning (p -value=0.019). The average posttest score for the treatment group is 28.5 (s.d. = 4.56), and the average posttest score for the control group is 27.3 (s.d. = 5.51). One thing to notice is that as displayed in Table 3, although in this matching, there were no significant differences between groups on any of the matching variables, the trend was always in favor of the control group, giving them an advantage. Nevertheless we still find the treatment group learned more.

The result here shows that for a pair of students who are categorized as similar learners by their prior knowledge and engagement level in the course, the one who displayed higher-order thinking behaviors had higher learning gains than the one who did not. We also computed the effect size of the treatment effect using the Cohen’s d method, which is 0.227 in this case.

4.3 Within-subject analysis over weeks

The second approach we adopted to account for individual difference is a within-subject analysis method. It is to investigate for each individual, whether the weeks he/she displayed more higher-order thinking behaviors would be associated with more learning than the other weeks for the same student.

4.3.1 Data preparation

In this analysis, we break down the students’ activities in the whole course into 12 weeks. And we use the time students submit each quiz to segment behavior into different weeks. For example, the activities the student did between the time he submitted quiz2 and quiz3 is considered the activities he did for week3. Using this method, each student will have one entry of aggregated data per week. We will introduce the variables we used as follows.

Variables that have a value per student per week:

QuizScore: This is the student’s quiz score of the week. In the case of the final week, this would be the post-test score. Typically each quiz consists of 10 questions related to the week’s material. The post-test consists of 25 questions for the whole course. Since the final quiz contains questions that do not belong to the material of that week, we dropped the final week data in our analysis for consistency. Thus, in our analysis, there are 11 weeks’ data per student.

Numpost: This is the number of posts the student contributed in that week.

For activities that come from the OLI website, it is possible to determine which content unit of the course each click is associated with. Two variables, namely *activities* and *pageviews*, are the number of activities and pages the student attempted on the topic of that week before he submitted the quiz of that week.

Activities: This is the number of activities on the OLI site the student attempted in that week.

Pageviews: This is the number of pages of OLI textbook the student viewed that week.

Playvideo: the number of videos the student played that week.

Variables that have a value per student:

Pre-test: As there are no pretests for each week, we used the pretest score of the course as a proxy for prior knowledge for each week.

All the above variables are standardized.

In order to compare the effect of higher-order thinking behaviors with that of paying general or focused attention to course materials, similar to the course-level analysis, we generated two binary group variables per student per week based on the highest level of discussion behavior the student displayed in the forum.

Group2[higher-order]: A binary variable, which equals to 1 if the student has contributed at least one constructive or interactive post during the week, otherwise it equals 0.

Group1[paying-attention]: A binary variable, which equals to 1 if the student has contributed at least one active post during the week but has not displayed any constructive or interactive post, otherwise it equals 0.

We sampled out students who have taken both pretest and posttest, and who have participated in the discussion forum at least once during the 11 weeks. Altogether, there are 404 such students used in our analysis. We then fitted a linear mixed-effect model using SPSS.

The major capabilities that differentiate mixed-effects models from general linear models are that mixed-effects models handle correlated data and unequal variances more effectively. In our case, as each individual has multiple entries that are considered to be correlated, we need to use a mixed-effect model.

In our model, we used students' QuizScore as a dependent variable, treated the student's CourseraID as a random effect, and treated all other variables as fixed effects. By doing this, we are accounting for participant effect, and observing the main effect of the factors related to each participant. The parameters of the mixed-effect model are shown in Table 5.

Table 5. Tests of Fixed Effects in the Mixed-Effect Model

Source	Numerator df	Denominator df	F
Group2 [higher-order]	1	4315.65	5.031*
Group1 [paying-attention]	1	4210.21	2.515
Intercept	1	434.261	21.111***
week	10	4101.57	37.367***
Pretest	1	398.322	19.844***
Activities	1	4137.68	148.022***
Pageview	1	4411.26	13.677***
Playvideo	1	4379.43	37.114***

(p<0.001***, p<0.01**, p<0.05*)

The result shows that when controlling for participant effect, for each individual, in the weeks he/she displayed higher-order thinking behaviors, he had better performance in the quiz relative to what was expected based on pretest score and other control variables than the weeks he did not display any higher-order thinking behaviors. However, in this model we do not see an effect of showing active attention to course materials.

Based on the results shown by propensity score matching, and now this within-subject analysis, we are more inclined to explain

the effect we see of higher-order thinking behaviors on learning as demonstrating that students are engaged in a more effective way with the course materials when they demonstrate these higher-order thinking behaviors, which leads to better learning.

5. WHICH TOPICS TRIGGER RICHER DISCUSSION?

From the analyses just shown, we see an effect of higher-order thinking behaviors on learning, and we also see that students' level of cognitive engagement varies from week to week. This drives us to ask a final research question, namely, which kind of learning materials trigger richer discussion?

In order to answer this question, we adopted a content analysis approach to extract topics from the textbook of the course and investigate which topics are associated with more high-order thinking behaviors.

5.1 Topic Modeling Setup

We used LDA (Latent Dirichlet Allocation) [4] topic modeling to extract topics from the OLI textbook. Latent Dirichlet Allocation (LDA) is a statistical generative model that can be used to discover hidden topics in documents as well as the words associated with each topic.

There are 16 chapters in the OLI textbook, although the last chapter is not used in the course. We trained a topic model using the LDA algorithm (provided in the Mallet package²) on the 15 chapters of the textbook. We split each chapter into 3 sentences units, and used it as the unit of analysis. We treated each unit as a document such that there are 2946 documents in total. The model was set to estimate 15 latent topics. We applied the trained topic model back to the documents and obtained a topic distribution over the 15 topics for each document. Each corpus unit is then represented as a vector of topics, as shown in (1). Corpus units represent the learning materials of each chapter of the textbook, we thus have a topic representation of the content of the learning materials.

$$Document = \sum_{i=1}^{15} w_i * T_i \quad (1)$$

Based on our annotation, we have counted the number of higher-order thinking behaviors students have displayed in each week. The course syllabus provides a mapping from 12 course weeks to the 15 chapters of the textbook. Based on this mapping, we could associate each document with the number of higher-order thinking behaviors students displayed in the corresponding week.

We then fitted a regression model using the corpus unit data, with the 15 topics as independent variables, and the number of higher-order thinking behaviors for each unit as a dependent variable, as shown in (2). By doing this, we can see which topics have a higher weight in predicting the number of higher-order thinking behaviors. Among the 15 topics, 1 was dropped in the regression model. For the rest of the topics, 5 are positively associated with richer discussion, 5 are negatively associated with richer discussion, and 3 do not have a significant effect.

$$Num(Higher\ order\ thinking) = \sum_{i=1}^{15} \beta_i * T_i \quad (2)$$

² <http://mallet.cs.umass.edu/topics.php>

5.2 Contrasting corpus units

We picked out two topics that have the highest positive weights, and two topics that have the highest negative weights as representatives for further analyses. The two topics that have the highest positive weights are “development” and “intelligence”. The two topics that have the highest negative weights are “neuroscience/brain” and “memory”.

We then looked into the OLI textbook to retrieve corpus units that have high and low weights on the positive and negative topics to explore the difference. In the topic model we built earlier, each corpus unit has a distribution over the 15 topics, for a given topic, e.g., “development”, we rank the corpus units based on their weight on this topic, and the units that have the highest weight or lowest weight are chosen as examples and displayed below.

We adopted a more qualitative approach to analyze the difference between these units. We present and contrast some illustrative examples below.

Based on the regression model we fitted, we consider that text units that have a higher weight on topic “development” and “intelligence” are associated with richer discussion, and that text units which have a lower weight on topic “development” and “intelligence” are associated with less cognitively rich discussion. We looked into text units in the chapter of development and intelligence, and found some examples in these two chapters that either have a high weight on these two topics or a low weight on these two topics. In the following example paragraphs, we use a symbol “+” to indicate that the text unit is positively associated with higher-order thinking behaviors; and we use a symbol “-” to indicate the text unit is negatively associated with higher-order thinking behaviors. For the positive examples, we underlined phrases that are daily life phenomenon; and for the negative examples, we underlined technical terms that we do not frequently use in daily life.

Here is an example of a unit that has a high weight on the topic of “development” in the development chapter:

(+) “During this stage children desire to experience pleasure through bowel movements, but they are also being toilet trained to delay this gratification. Freud believed that if this toilet training was either too harsh or too lenient, children would become fixated in the anal stage and become likely to regress to this stage under stress as adults. If the child received too little anal gratification (i.e., if the parents had been very harsh about toilet training), the adult personality will be anal retentive—stingy, with a compulsive seeking of order and tidiness.”

In the same chapter, here is a unit that has a low weight on the topic of “development”:

(-) “The medical research was a 1998 study published in the prestigious medical journal, *The Lancet*, by a British physician named Andrew Wakefield. He and his colleagues reported data allegedly collected from twelve children who had diagnoses of regressive autism, 11 of whom also had a diagnosis of non-specific colitis. Wakefield’s paper claimed that this new brain and bowel disease “syndrome” (autistic enterocolitis was the term coined by Wakefield for the paper) started very soon after administration of MMR vaccine, as reported to Wakefield’s team by the parents of children in the study.”

Here is an example of a unit that has a high weight on the topic of “intelligence” in the intelligence chapter:

(+) “Once the standardization has been accomplished, we have a picture of the average abilities of people at different ages and can

calculate a person’s mental age, which is the age at which a person is performing intellectually. If we compare the mental age of a person to the person’s chronological age, the result is the intelligence quotient (IQ), a measure of intelligence that is adjusted for age.”

In the same chapter, here is a unit that has a low weight on the topic of “intelligence”:

(-) “Severe and profound mental retardation is usually caused by genetic mutations or accidents during birth, whereas mild forms have both genetic and environmental influences. One cause of mental retardation is Down syndrome, a chromosomal disorder leading to mental retardation caused by the presence of all or part of an extra 21st chromosome. The incidence of Down syndrome is estimated at 1 per 800 to 1,000 births, although its prevalence rises sharply in those born to older mothers.”

We also extracted some units that have high weight on the topics “brain” and “memory” for comparison. We consider the units that have a high weight on the topics “brain” and “memory” are negatively associated with rich discussion. Here are two examples that have a high topic representation of “brain” and “memory” respectively.

(-) “As you can see in the following figure, neurons consist of three major parts: a cell body, or soma, which contains the nucleus of the cell and keeps the cell alive; a branching, treelike fiber known as the dendrite, which collects information from other cells and sends the information to the soma; and a long, segmented fiber known as the axon, which transmits information away from the cell body toward other neurons or to the muscles and glands. Some neurons have hundreds or even thousands of dendrites, and these dendrites may be branched to allow the cell to receive information from thousands of other cells. The axons are also specialized, and some, such as those that send messages from the spinal cord to the muscles in the hands or feet, may be very long—even up to several feet in length.”

(-) “Short-term memory (STM) is the place where small amounts of information can be temporarily kept for more than a few seconds but usually for less than one minute. The cognitive psychologist George Miller referred to “seven plus or minus two” pieces of information as the “magic number” in short-term memory. Information in short-term memory is not stored permanently but rather becomes available for us to process, and the processes that we use to make sense of, modify, interpret, and store information in STM are known as working memory.”

In the two positive examples in the chapter of development and intelligence, the content is explaining life phenomenon that are familiar, for example, toilet training or IQ tests. And there are no technical terms in the text. All words are what we use in daily life. By contrast, even in the same chapter, the two examples that have a lower weight over the topics “development” and “intelligence” include more technical/medical terms that we seldom use in daily life. It is the same case for the two examples on the topic of “brain” and “memory”. The two examples are about biopsychological processes that we seldom experience in life. And there are a lot of unfamiliar technical terms in the text, making it more challenging to read.

In the following table, we display the words that are associated with topics that have a strong positive or negative association with the prevalence of higher-order thinking behaviors students displayed.

Table 6. Words associated with positive and negative topics

Positive	Negative
Intelligence; People; Children; Development; Parents; Adults; Relationships; Social; Language; Group; Personality; Behavior; Arousal; Experience; Emotions; Women; Body; System	Memory; information; brain; neurons; cortex; cells; nerve; hemisphere; disorders; psychological; mental; symptoms; visual; sound; eye; perceptual;

By comparing and contrasting the words associated with positive and negative topics, we can see that more social oriented topics are positively associated with richer discussion. These examples show that these topics use more daily life words and explain human phenomenon people are more familiar with. On the other hand, more neuroscience oriented topics are negatively associated with richer discussion, and the examples show that these topics use more technical terms that are not usually used in daily life. They also explain animal experiments or biopsychology phenomenon that are not commonly familiar.

6. DESIGN IMPLICATIONS

Based on our findings and discussion, we propose the following design implications, which could be a next-step intervention built into MOOC discussion forums to support richer discussion.

- 1) We see that content related to daily life and social experience is associated with higher-order thinking behaviors. As explained in motivation theories [10,15], real-world connection helps support situational interest which could lead to higher intrinsic motivation in the student. For the units that are not connected to personal interests and experiences, e.g., neuroscience, we can think of manipulations to increase student motivation. Some practices may include, incorporating blocks of real-life stories in the unit, or designing collaborative tasks that require students to connect the concepts to personal life.
- 2) We see that the contents that have more technical terms are associated with cognitively poorer discussion. We explained this phenomenon as people tend to discuss less when they understand less. This may suggest that support for reading comprehension where technical terms come up might improve productive engagement with the material. Further, scaffolding for explanation should be provided in the units that have less of a tendency to elicit explanation behavior.
- 3) We also found that a lot of discussions in the forums are off-topic. One way to address this is to take the off-topic conversation that comes and try to channel it towards course content. As it stands, off-topic conversation may not be valuable for learning, but if we can use it to draw attention to personal connections with the course material, it could become valuable. Providing scripts [13] to achieve this could also be a future direction.
- 4) In this paper, we developed a coding manual to categorize students' discussion behaviors by the cognitive engagement displayed in their discussion. In a next step, we want to build machine learning models to automatically label students' discussion behaviors as in our prior work [26], thus enabling the provision of just-in-time feedback. For example, if the analysis reveals that a student is constantly engaging in off-topic discussion in the forum, we could respond by providing scripts to direct him towards course content. Or if we see a student is constantly repeating or paraphrasing course materials without constructing ideas on top of that, we could also provide scripts for these students, e.g, asking probing questions to get them think about the content they are posting.

7. CONCLUSION AND LIMITATION

In order to better support discussion in a MOOC context, we investigate what kinds of discussion behaviors are associated with learning and what types of learning materials trigger richer discussion. We developed a coding manual based on Chi's [8] ICAP framework and situated it in a MOOC context to categorize students' posts in discussion forums based on different observed levels of cognitive engagement. We achieved high reliability when applying this coding manual in a psychology MOOC.

Driven by the first research question, we found that students who displayed more higher-order thinking behaviors learnt more through deeper engagement with course materials displayed by their discussion behaviors. In this course, students who displayed higher-order thinking behaviors learnt more than students who simply directed their attention to course materials. These students in turn also learnt more than students who were constantly off-topic in the forums.

The follow-up analysis with LDA topic modeling applied to course materials informs us that social oriented topics triggered richer discussion compared with biopsychology oriented topics, and that higher-order thinking behaviors tend to appear together within threads in the forums. We have suggested design recommendations based on these observations.

However, there are limitations in this work, which could be addressed and improved in future research.

- 1) There are relatively few instances of interactive behaviors in the dataset, so we grouped together constructive and interactive behaviors as higher-order thinking behaviors. Because of this, we are not able to compare the effect between interactive behaviors and constructive behaviors on learning. A next step is to apply this coding manual to additional courses so that we will have more data to address this sparsity issue. In doing this, we will also have more training data to build machine learning models to detect the discussion behaviors automatically.
- 2) In our approach, we are only able to estimate students' cognitive engagement if they display it by posting in the discussion forum. It is possible that students are highly engaged with course materials even if they never display those thinking behaviors in the discussion forum. Although we have already controlled for students' engagement in other course activities, we are losing information about invisible learning traces. Discussion data would be more valuable as a lens on engagement if more students posted to the forum. In our future research through deploying interventions, we aim to engage more students in the forums so that assessment based on discussion behavior can be applied to a higher proportion of students in the course.

8. ACKNOWLEDGEMENTS

This project is funded by NSF Grant ACI-1443068 and funding from Google.

9. REFERENCES

- [1] Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Are you reading my mind?: modeling students' reading comprehension skills with natural language processing techniques. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM. DOI= <http://dx.doi.org/10.1145/2723576.2723617>.
- [2] Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web* (pp. 687-698). International World Wide Web

- Conferences Steering Committee. DOI=
<http://dx.doi.org/10.1145/2566486.2568042>
- [3] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3), 399-424. DOI=
<http://dx.doi.org/10.1080/00273171.2011.568786>
- [4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.
- [5] Brinton, C. G., Chiang, M., Jain, S., Lam, H. K., Liu, Z., & Wong, F. M. F. (2014). Learning about social learning in MOOCs: From statistical analysis to generative model. *Learning Technologies, IEEE Transactions on*, 7(4), 346-359. DOI=
<http://dx.doi.org/10.1109/TLT.2014.2337900>.
- [6] Brooks, C., Chavez, O., Tritz, J., & Teasley, S. (2015). Reducing selection bias in quasi-experimental educational studies. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 295-299). ACM. DOI=
<http://dx.doi.org/10.1145/2723576.2723614>
- [7] Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471-533. DOI=
[http://dx.doi.org/10.1016/S0364-0213\(01\)00044-1](http://dx.doi.org/10.1016/S0364-0213(01)00044-1)
- [8] Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219-243. DOI=
<http://dx.doi.org/10.1080/00461520.2014.965823>
- [9] Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research*, 64, 1-35. DOI=
<http://dx.doi.org/10.3102/00346543064001001>
- [10] Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology*, 99(3), 597. DOI=
<http://dx.doi.org/10.1037/0022-0663.99.3.597>
- [11] Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 146-150). ACM. DOI=
<http://dx.doi.org/10.1145/2723576.2723589>
- [12] Ferguson, R., & Clow, D. (2015). Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM. DOI=
<http://dx.doi.org/10.1145/2723576.2723606>
- [13] Fischer, F., Kollar, I., Stegmann, K., Wecker, C. & Zottmann, J. (2013). Collaboration scripts in computer-supported collaborative learning. *The international handbook of collaborative learning*, 403-419.
- [14] Greer, J., & Mark, M. (2015). Evaluation Methods for Intelligent Tutoring Systems Revisited. *International Journal of Artificial Intelligence in Education*, 1-6.
- [15] Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational psychologist*, 41(2), 111-127. DOI=
http://dx.doi.org/10.1207/s15326985ep4102_4
- [16] Hsiao, I. H., & Awasthi, P. (2015). Topic facet modeling: semantic visual analytics for online discussion forums. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 231-235). ACM. DOI=
<http://dx.doi.org/10.1145/2723576.2723613>
- [17] Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170-179). ACM. DOI=
<http://dx.doi.org/10.1145/2460296.2460330>.
- [18] Klusener, M., & Fortenbacher, A. (2015). Predicting students' success based on forum activities in MOOCs. In *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, 2015. IEEE. DOI=
<http://dx.doi.org/10.1109/IDAACS.2015.7341439>
- [19] Koedinger, K. R., Kim, J., Jia, J. Z., McLaughlin, E. A., & Bier, N. L. (2015, March). Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 111-120). ACM. DOI=
<http://dx.doi.org/10.1145/2724660.2724681>
- [20] Koedinger, K. R., McLaughlin, E. A., Jia, J. Z., & Bier, N. L. (2016). Is the Doer Effect a Causal Relationship? How Can We Tell and Why It's Important. In *Proceedings of the 6th International Conference on Learning, Analytics and Knowledge*.
- [21] McKendree, J., Stenning, K., Mayes, T., Lee, J., & Cox, R. (1998). Why observing a dialogue may benefit learning. *Journal of Computer Assisted Learning*, 14(1), 110-119. DOI=
<http://dx.doi.org/10.1046/j.1365-2729.1998.1420110.x>
- [22] McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, 14(1), 1-43. DOI=
http://dx.doi.org/10.1207/s1532690xci1401_1
- [23] Robinson, A. (2015). Exploring Class Discussions from a Massive Open Online Course (MOOC) on Cartography. In *J. Brus et al. (eds.), Modern Trends in Cartography, Lecture Notes in Geoinformation and Cartography*. DOI=
http://dx.doi.org/10.1007/978-3-319-07926-4_14
- [24] Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., & Sherer, J. (2014, March). Social factors that contribute to attrition in moocs. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM. DOI=
<http://dx.doi.org/10.1145/2556325.2567879>
- [25] Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33-38. DOI=
<http://dx.doi.org/10.1080/00031305.1985.10479383>
- [26] Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating how student's cognitive behaviors in MOOC discussion forums affect learning gains. In *Proceedings of the 8th International Conference on Educational Data Mining*.
- [27] Zhu, E. (2006). Interaction and cognitive engagement: An analysis of four asynchronous online discussions. *Instructional Science*, 34(6), 451-480.