Final Project

Tanha Kate

FE50 Fall 2017

Research Question:

How can we systematically and routinely use data from the Active Learning Forum to measure student progress?

1.Introduction

Minerva Schools at KGI, hereafter referred to as Minerva, is an innovative university that provides a rigorous interdisciplinary study, global immersion in seven different countries, 'merit-based, need-blind admissions' and a diverse student body. (Minerva Project, 2017) Minerva's modus operandi in executing its educational model is the Active Learning Forum (ALF), an online software platform designed to engage students through Socratic discussion in live video interactions. The magnitude of data gathered on the ALF is enormous in terms of the observations per student and types of information, such as student demographics, clickstream, text-based submission, and context-specific data about video responses and poll answers. In this work, the author analyzes multiple educational variables to explore the routine use of data in operationalizing student progress. Note that due to the Family Educational Rights and Privacy Act (FERPA), the analysis is limited to a sample size of one student. The key findings of this paper are that session engagement shows no correlation with high grades, individual engagement is not correlated with overall peer engagement and that modeling outcome grades over time shows a positive trend.[1]

[1] **#thesis:** I have used a front-door approach to state my thesis statement. It succinctly states the results of my paper and takes a data-driven, arguable stance. I also defined the scope of my paper later on by specifying sample size, semester duration and specific course.

2. Review of Literature

The standard way of thinking about education innovation has it that ‘active, engaged minds learn much better than students who passively listen to lectures’ (Katzman et al., 2017). In fact, a meta-analysis of undergraduate science, technology, engineering, and mathematics (STEM) courses show that 'students in classes with traditional lecturing are 1.5 times more likely to fail' than students engaged in active learning. (Freeman et al., 2014) Active learning is ‘any instructional technique which requires students to apply or process content as part of the learning experience’. (Powner et al., 2008) Inefficient instructional techniques in the undergraduate classroom have an especially detrimental effect on women and other minority groups: a study of 23 large introductory biology classes found that women made up 'less than 40% of those heard responding to instructor-posed questions', despite a numerical dominance of women in such classrooms. (Eddy et al., 2014) Therefore, performance metrics for female students, ‘regardless of academic standing’, are negatively impacted by classroom anxiety. (Ballen, 2017)

One of the significant divergences from the standard educational pedagogy of traditional lecture-based classrooms, came in the form of edX - a jointly founded platform by HarvardX and MITx for delivering massive open online courses (MOOCs). With the goals of ‘increased access to educational opportunities worldwide, enhancement of on-campus education, and research about effective technology-mediated education’, edX was launched in 2012 and enlisted 597,692 unique users, with varying genders, levels of education completion and nationalities across 17 different courses. (Ho et al., 2014) The larger number of registrants, low cost, time flexibility, and asynchronous use of course material (resources remain accessible even after the duration is complete and enrollment rises when certification is not possible. (Ho et al., 2014)) are the major structural differences between edX and the ALF's offerings of the first-year undergraduate courses, namely the Cornerstones.[2] However, both edX and the ALF are rapidly evolving online platforms and the former's operationalization of educational variables like 'enrollment,' and 'achievement' can help abstract key indicators of student engagement.

Certification on the edX platform is a 'poor proxy for learning' and demonstrates a' short-sighted, misleading metric for engagement'. (Ho et al., 2015, Seaton, 2014) Figure 1 plots distributions of the percent of course chapters accessed by students along the horizontal axis. Note that vast numbers of students exist in all regions of the scatterplot. Because attaining a certificate is only one possible learning pathway, completion rates and a comparison of grades do not encapsulate the diversity of learning approaches that registrants take in virtual environments.

---

[2] **#analogies**: I identified deep structural similarities and dissimilarities in terms of learning and environment features on edX, a platform analogous to the ALF and reverse engineered techniques for variable selection and operationalization.

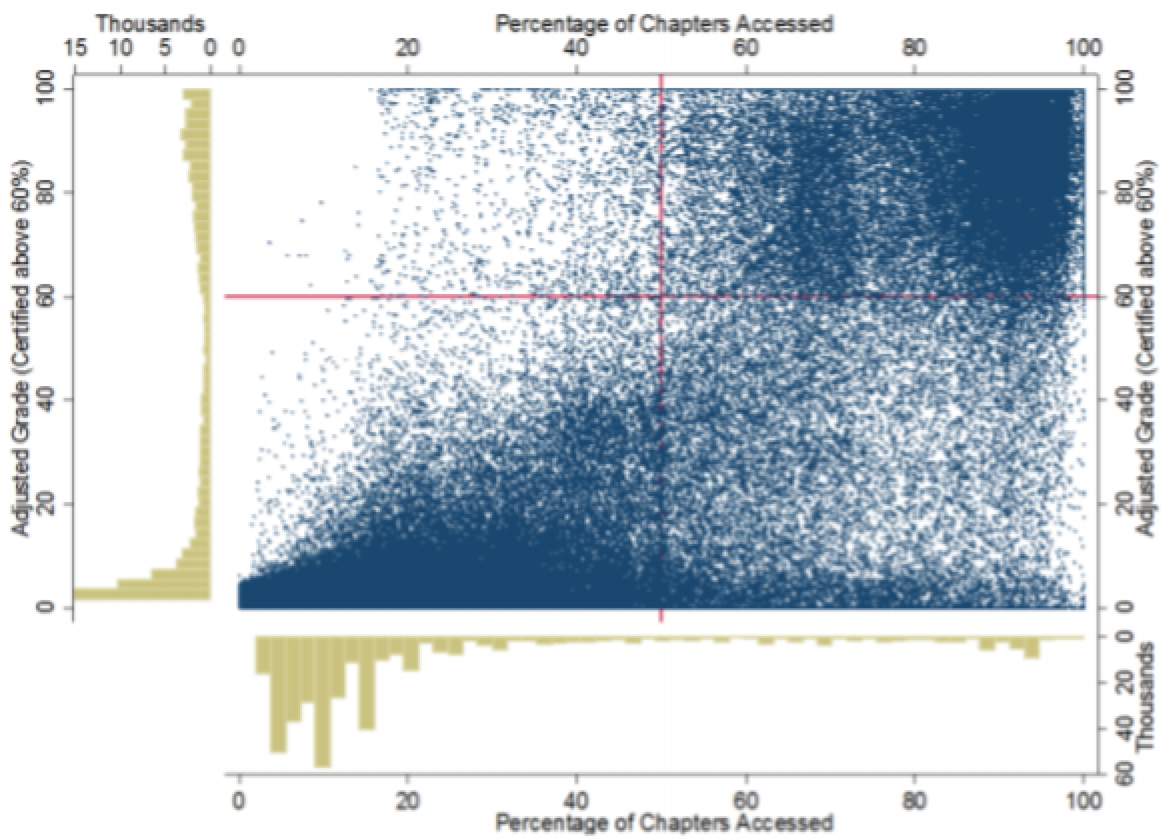


Figure 1. Distributions of course activity (in terms of the percentage of chapters accessed) and course grades (for grades above 1%, linearly adjusted across courses to a common certification cutoff of 60%) (Source: Ho et al., 2014)

Other variables, which are rare in on-campus college courses but offer avenues for rigorous research include 'socioeconomic status, motivations for partaking in particular courses, detailed video interaction behaviors, number of clicks and number of active days in the course'. (Ho et al., 2014) However, since these variables are often context-dependent, careful selection is particularly important because it informs the course design and directs further research into effective learning principles both online and offline. For instance, studies investigating the correlation between time spent on specific course resources and various measures of student performance in MOOCs show weak or slightly negative correlations (Champaign et al., 2014).

This is aligned with findings about the negligible correlation between the amount of studying and grades, disproving assumptions that 'college grades reflect student effort.' (Schuman et al., 1985)

Minerva's technology platform synthesizes outstanding, cognitively undemanding features with advances into the effectivity of science of learning and peer instruction. (Katzman et al., 2017) Furthermore, the dynamism of live-streaming introduces a 'social component,' which has proven important in motivating students to engage in learning materials collaboratively. (Bader-Natal, 2009 & Mason et al., 2013) The benefits of peer instruction extend to both 'higher and lower background-knowledge students' (Lasry et al., 2008) and do not necessitate anyone in the group initially knowing the correct answer. (Smith et al., 2009). Further features of the platform that enables effective discussion are summarized below:

Criteria	**Feature**
Majority class participation	Facilitated by professors calling upon students depending on decision support tools which display the duration for which students have spoken (Katzman et al., 2017)
Dialogue among students with varying personality preferences, i.e., extraversion and introversion	Facilitated by smaller break out groups with fewer participants and large group socratic sprints (Katzman et al., 2017), which refine understanding, review main points, clear up misconceptions and expand student interest
Exposes participations to diverse perspectives which encourage the comparison of ideas and formation of 'atypical combinations' new ideas (Uzzi et al., 2007)	Minerva's student body diversity accomplishes better than any other institution by its selection process, which extends opportunities to students from all over the world and across socioeconomic backgrounds.
Students are prepared and have the same background knowledge	Facilitated by assigned readings for each class and graded pre-class work

Students at Minerva are graded on their responses in classes and assignments against 'a core set of Habits of Mind and Foundational Concepts (HCs)' on a 5-point grading scale. (Katzman et al., 2017) Within a Minerva Cornerstone class, course structure constitutes a few base sources (reading materials, videos, academic papers), each categorized under a specific session which explores an aspect of a more substantial question with associated HC's. Class sequences form the bulk of learning in each session. Each sequence consists of an introduction to course objectives, smaller breakout activities interspersed with seminar-style Socratic sprints. Graded poll questions are administered at the beginning and end of the session, and after a portion of the content is covered. The ALF allows faculty to 'review and directly assess students' in-class contributions' at the 'granularity of an individual spoken comment or a typed comment in chat', anchored to a specific moment in video or written assignments/code. (Levitt et al., 2017) Minerva's 'HC assessment dashboard' presents these scores in chronological order, providing both feedback and assessment in context. (Levitt et al., 2017) In summary, the ALF incorporates a multitude of open discussion formats that encourage peer-instruction and positively reinforce participation that yield data for systematic assessment of students across multiple levels of analysis against specific desired outcomes.

3. Methodology

3.1 Study sample

Because the data and assessment tools for the higher year courses are at a nascent stage of development (as well as FERPA restrictions), this paper will focus exclusively on Cornerstone Classes. The following visualizations use real data from the Formal Analysis course at Minerva, hereafter referred to as FA50, unless specified otherwise. The author does not intend this to be a full review of the ALF's data structures or courses and uses FA50 to illustrate general trends. Since the course content is viewed as challenging, the author concludes that an extreme case will help abstract trends that are generalizable across more massive datasets. FA50 focuses on thinking critically: students are trained in advanced logic, rational thought, introductory programming, statistics and computational thinking. The data is collected from the beginning of the 2017 Fall semester and runs a duration of roughly 4.5 months. During the data cleaning process, sessions that haven't been graded were excluded (n=20).

An important consideration is the sample size (n=1) which limits the reliability of conclusions. Furthermore, research shows that students' 'prior experience with similar pedagogy' is a structurally significant factor influencing the extent to which they benefit. (Zhao et al., 2014) The subject, who previously studied in similar seminar-style classrooms, was confident in her ability to improve through Minerva's active learning practices. The author acknowledges the effect of this inherent bias in any conclusive results from the study.

3.2 Operationalizing Variables and Procedure

The ALF offers diverse, complete logs of activities for each student. The paper's major objective is to investigate the association between the subject's talktime in each session, i.e., how long the subject has spoken during the entire duration of class (excluding breakout groups) and the score on text/video responses in that session as a measure of student achievement.

Engagement is operationalized through talk time for the following reasons:

a) Majority of assessments are conducted in class, where the primary means of interaction is verbal. (Figure 2) Assignments often have higher weights than classroom assessments on the overall cumulative score. However, the dataset deals exclusively with unweighted scores, and association will not include assignment data which shows highest average scores for the subject. (see table below)

Types of Assessments	**Mean Score**
Assignment	3.50
Video	3.27
Poll	3.16

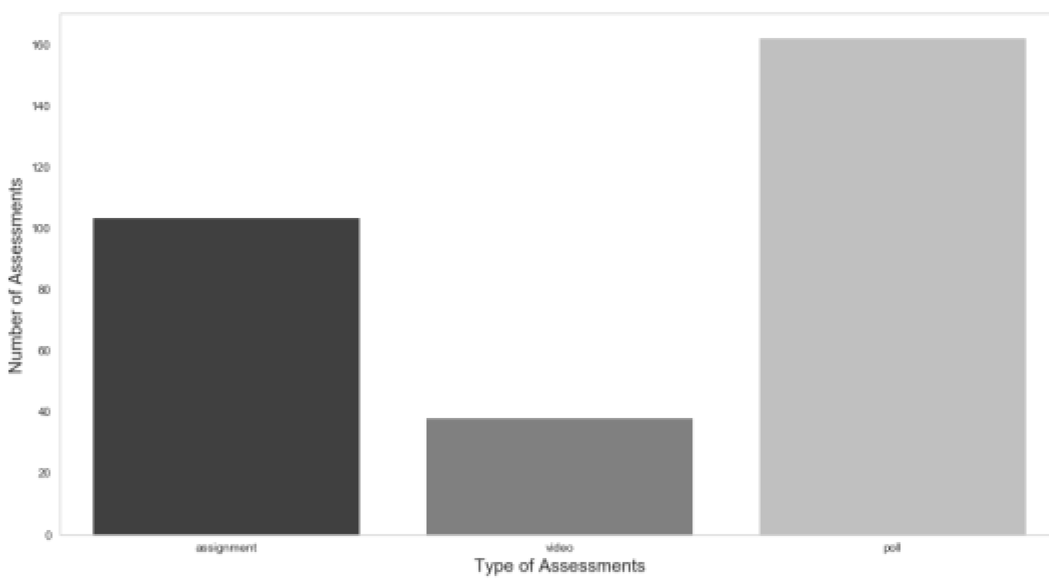


Figure 2. Distribution of types of assessments in Minerva across all first-semester courses. The graph shows that video and poll assessments constitute the majority of assessments. (n=334)

b) It is an easily quantifiable variable of social interaction in class, which is identified earlier as one of ALF's significant strengths as a learning platform. This assumes that all audio input in the session was related to course content, including responses that expressed inability to answer questions or unmindful enabling of talking features.

c) Recent studies show that 'people's use of each online platform was significantly related to engagement.' (Paek et al., 2013) Since the ALF offers video-streaming functionality, talk time can serve as a proxy data for its use. This assumption disadvantages students who speak less in class but remain attentive to the contribution of others', chatroom discussions and take detailed notes. [3]

Since talktime by itself leaves out important contextual information about the overall level of participation of all students in class, deviation from the class average for the subject's talk time will be considered. The class average also excludes the subject's contribution and breakout session talktimes for other students. Achievement is operationalized as HC scores, supporting interpretations about student skill level with respect to the course's learning objectives.

[3] **#evidencebased**: I have compiled evidence from the various academic readings, explained the substantiated argument and elaborated on its relevance to the subject matter of my paper. To do this successfully, I have explicitly stated an example of how engagement was measured in similar learning platforms to strengthen my argument.

3.3 Major Hypothesis[4]

Null Hypothesis (H_0):	Higher talk time has no effect on students' achievement.
Alternative Hypothesis (H_A):	Higher talk time has a positive effect on students' achievement.

Findings

1. Engagement shows no correlation with higher scores.

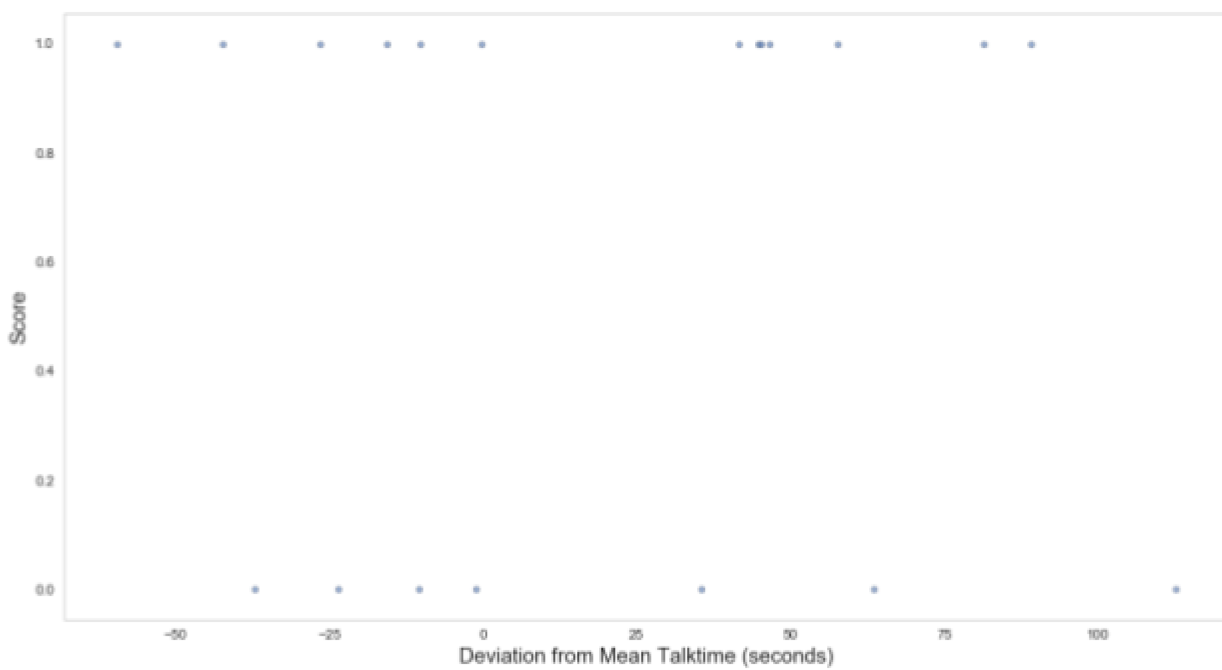


Figure 4. Scatterplot displaying deviation of subject's total talktime from mean talktime (seconds) against scores. Since subjects session score data for FA50 only assumed values of 3 and 4, they have been coded into 0 and 1 respectively. The value of Spearman's rho is 0.02727 and the two-tailed value of P is 0.90914.

---

[4] **#hypothesisdriven:** From accepted scientific knowledge about the science of learning and educational pedagogy, I have inductively generate a specific hypothesis. My hypothesis is both explanatory in terms of the association between engagement and achievement and predictive in terms of the possible nature of this association.

Since HC scores are discrete and ordinal, we used Spearman's Rank Correlation. By normal standards, the association between the variables would not be considered statistically significant. Therefore, even though the results indicate a weak positive correlation, the author failed to reject the null hypothesis (H_0). These findings may have implications on current designs of active learning, where instructors at the college level dedicate large portions of class time to class participation - which is often perceived by instructors as a sign of intelligence and can therefore influence their behavior towards the student. (Smith et al., 2013) Further analysis of talk times and assessment data from larger, representative samples of Minerva students can provide insights on the association between engagement and achievement.

2. Individual engagement is not correlated with overall peer engagement

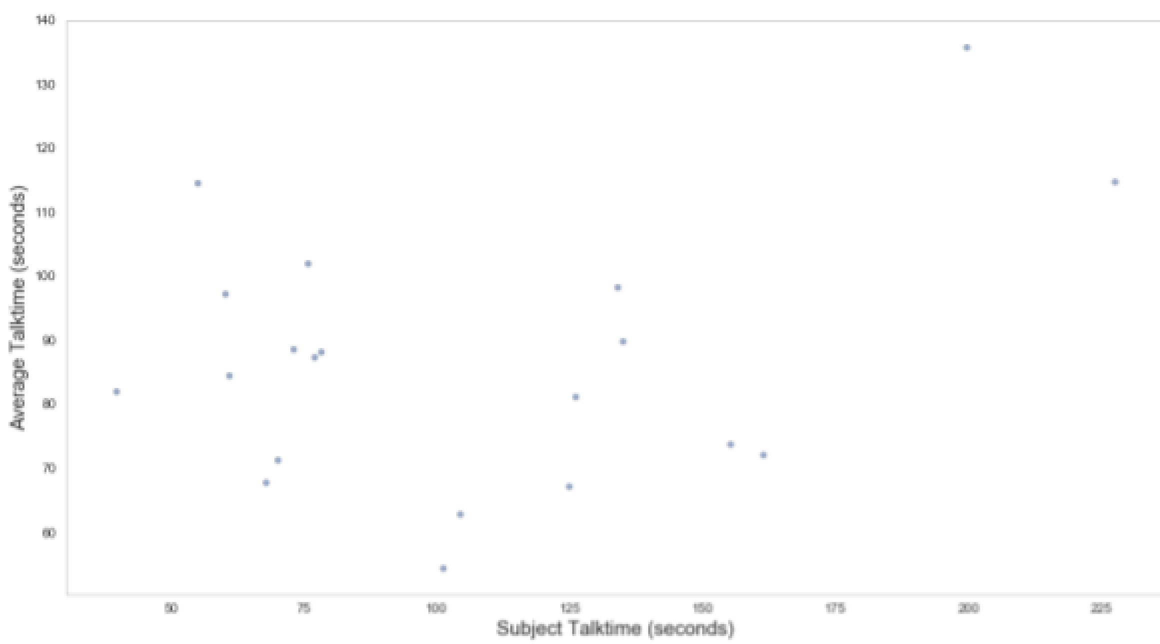


Figure 3. Scatterplot displaying subject talktime against average talktime. All values are recorded in seconds. The value of Pearson's r is 0.3075.

Although a positive Pearson's correlation, the relationship between the variables is weak. The author therefore concludes that average talk time for all students and the subject's talktime do not affect each other.

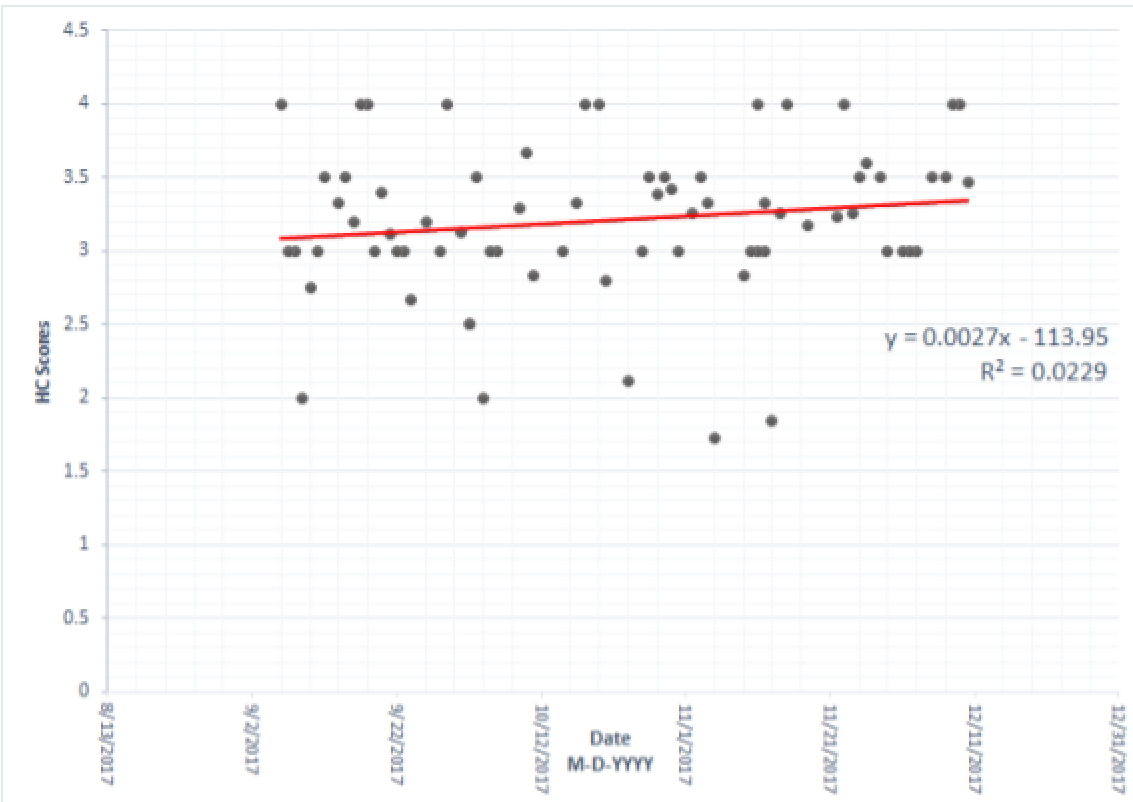
However, statistical analysis of small world network shows that 'open networks with multiple links between people from different clusters' lead to greater relative performances. (Simmons, 2015) On the ALF, Minerva students have such a network, where each participant, coming from vastly different social systems, have the ability to share their experiences and knowledge. Multiple interaction formats can expose individual students to information from diverse clusters and create intellectual safety, especially in an assessment style that incentivizes participation.[5] (McLoughlin et al., 2010) However, there are many confounding variables[6] which affect the classroom experience including instructor gender, number of students, whether the course is optional or required. (Eddy et al., 2014) With a larger dataset, it will be possible to make more substantial remarks about the correlation between increased overall engagement and individual engagement. This will in turn provide data-driven insights into the efficacy of Minerva's classrooms.

[5] **#emergentproperties**: This is an example of emergent properties because I have shown how the heterogeneity of student population in Minerva and the diversity of opinions lead to a feeling of intellectual safety. Furthermore, it is an effective application of appropriate **#levelsofanalysis** because I have honed in on both individual data as well as group level interactions on the ALF.

[6] **#variables:** Throughout this paper, I have operationalized variables purposefully. This is an effective application of the HC because it also incorporates an understanding of confounding variables which often have an extraneous effect on the study.

<u>**3.** Modelling outcome grades over time shows a positive trend.</u>

To develop a proxy measure for the subject's relative improvement in the course, the author fits a simple linear regression line for the subject's scores over the duration of the first semester and uses the slope of this line as the student's relative improvement. (Figure 5)



[7]

Figure 5. Linear regression model for mean HC scores against date.

---

[7] **#dataviz:** This presents effective data visualization techniques because I have manipulated the raw data to extract meaningful insights regarding trends in my outcome scores. I have also critiqued on the limitation of this visualization in terms of small sample size and low R^2. This is also an effective application of **#algorithms** because I've used a combination of Pythonic and C++ code to clean up my data and evaluate HC means for each date assessed and create clear data visualizations. This displays my ability to break down a problem effectively across multiple languages. [see appendix]

Since the gradient of the regression line is positive, it is plausible that the subject's scores have risen over the course of the semester. However, a low R^2 value indicates that data is not fitted around the regression line. Furthermore, linear regression assumes that the relationship between the independent and dependent variables are linear. This assumption is tested with the scatter plot, which displays little linearity, few outlier values and heteroscedasticity (residuals are not equal across the regression line). Our model is further limited by the assumption that prior to starting the semester, there are no HC scores for the subject i.e. the subject had no academic achievement - which is unlikely considering Minerva's highly selective admissions process.

4. Conclusion

This study has found no correlation between talktime and achievement, individual engagement and overall engagement, and a positive trend in HC scores across the semester. The larger sample size and the kind of educational data on the ALF offers insurmountable opportunities to further observe, analyze, and ultimately improve the learning process by focusing on both cumulative student data as well as the specific trajectory of a single student. By monitoring the effects of a new pedagogy on the highly diverse cohorts that undergo the Cornerstone courses annually, it is possible to perform various data mining and learning experiments. This can be achieved with some combination of generalization/suppression or under a 'framework of differential privacy' (Daries et al., 2014) that protects personally identifiable information. Randomized representative samples of the student population will lend tremendous support to interpretations about the causal impact of the Minerva's active learning model, illuminate the best metrics of student progress, and specify where to direct resources to improve student performance.

References[8]

Bader-Natal, A. Interaction Synchronicity in Web-based Collaborative Learning Systems. Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2009 (E-LEARN 2009), pp. 1121-1129.

Ballen, C., Salehi, S., & Cotner, S. (2017). Exams disadvantage women in introductory biology. *Plos One, 12*(10).

Champaign, J., Colvin, K. F., Liu, A., Fredericks, C., Seaton, D., & Pritchard, D. E. (2014). Correlating skill and improvement in 2 MOOCs with a student's time on tasks. Proceedings of the First ACM conference on Learning @ scale conference - L@S '14 , 11-20.

E., B., & W. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. Cbe Life Sciences Education, 13(3), 478-92. doi:10.1187/cbe.13-10-0204

[8] **#presentation:** I have followed the conventions of academic writing and cited all my sources appropriately in APA format. Furthermore, my paper draws on high #sourcequality academic papers across multiple disciplines to support my arguments and analysis.

F., E., M., S., O., J., & W. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences of the United States of America,111(23), 8410-5. doi:10.1073/pnas.1319030111

Ho, Andrew Dean and Chuang, Isaac and Reich, Justin and Coleman, Cody Austun and Whitehill, Jacob and Northcutt, Curtis G and Williams, Joseph Jay and Hansen, John D and Lopez, Glenn and Petersen, Rebecca, HarvardX and MITx: Two Years of Open Online Courses Fall 2012-Summer 2014 (March 30, 2015). Available at SSRN: https://ssrn.com/abstract=2586847 or http://dx.doi.org/10.2139/ssrn.2586847

Ho, Andrew Dean and Reich, Justin and Nesterko, Sergiy O and Seaton, Daniel Thomas and Mullaney, Tommy and Waldo, Jim and Chuang, Isaac, HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013 (January 21, 2014). Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1).. Available at SSRN: https://ssrn.com/abstract=2381263 or http://dx.doi.org/10.2139/ssrn.2381263

Katzman, J., Regan, M., and Bader-Natal, A. The Active Learning Forum. In Building the Intentional University: Minerva and the Future of Higher Education, edited by Stephen M. Kosslyn and Ben Nelson, MIT Press, 2017.

Lasry, N., Mazur, E., & Watkins, J. (2008). Peer instruction: From harvard to the two-year college. American Journal of Physics, 76(11).

Levitt, R., Bader-Natal, A., and Chandler, V. Assessing Student Learning. In Building the Intentional University: Minerva and the Future of Higher Education, edited by Stephen M. Kosslyn and Ben Nelson, MIT Press, 2017.

Mason, G. S., Shuman, T. R., & Cook, K. E. (2013). Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-‐division engineering course. IEEE Transactions on Education. Retrieved August 25, 2013, from http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6481483&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6481483

McLoughlin, C., & Lee, M. (2010). Developing an online community to promote engagement and professional learning for pre-Service teachers using social software tools. *Journal of Cases on Information Technology, 12*(1), 17-30. doi:10.4018/jcit.2010010102

Minerva Project. (2017). Retrieved June 22, 2017 from https://www.minerva.kgi.edu/

Minerva Schools at KGI. (2017). Outcome Assessments [JSON file]. Unpublished raw data. Retrieved from https://seminar.minerva.kgi.edu/api/v1/outcome-assessments

Minerva Schools at KGI. (2017). Talktime (Cornerstone Courses). Unpublished raw data. Retrieved from https://seminar.minerva.kgi.edu/api/v1/outcome-assessments

Paek, H., Hove, T., Jung, Y., & Cole, R. (2013). Engagement across three social media platforms: An exploratory study of a cause-related pR campaign. Public Relations Review, 39(5), 526-533. doi:10.1016/j.pubrev.2013.09.013

Powner, L., & Allendoerfer, M. (2008). Evaluating hypotheses about active learning. International Studies Perspectives, 9(1), 75-89.

S., W., A., W., K., G., & S. (2009). Why peer discussion improves student performance on in-class concept questions. Science (new York, N.y.), 323(5910), 122-4. doi:10.1126/science.1165919

Schuman, H., Walsh, E., Olson, C., & Etheridge, B. (1985). Effort and reward: The assumption that college grades are affected by quantity of study. Social Forces, 63(4), 945-966.

Seaton, Daniel Thomas and Coleman, Cody Austun and Daries, Jon P. and Chuang, Isaac, Teacher Enrollment in MITx MOOCs: Are We Educating Educators? (October 27, 2014). http://www.educause.edu/ero. Available at SSRN: https://ssrn.com/abstract=2515385 or http://dx.doi.org/10.2139/ssrn.2515385

Simmons, M. (2015, January 15) The no. 1 predictor of career success according to network science.*Forbes*. Retrieved from https://www.forbes.com/forbes/welcome/?toURL=https://www.forbes.com/sites/michaelsimmons/2015/01/15/this-is-the-1-predictor-of-career-success-according-to-network-science/&inf_contact_key=a86c92d8b745c1b6746de9a8fe9bb7d30934ee648760fe6d5f4ce82cd7521314&refURL=&referrer=#7ad9d7593623

S., J., G., & W. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *Cbe Life Sciences Education, 12*(4), 618-27. doi:10.1187/cbe.13-08-0154

Sweeney, L. Datafly: A system for providing anonymity in medical data. Database Security, XI: Status and Prospects. T. Lin and S. Qian, eds. Elsevier Science, Amsterdam. 1998

Uzzi, B., Amaral, L., & Reed-Tsochas, F. (2007). Small-world networks and management science research: A review. European Management Review, 4(2), 77-91. doi:10.1057/palgrave.emr.1500078

Zhao, Y., & Ho, A. (2014). Evaluating the Flipped Classroom in An Undergraduate History Course. Manuscript.

Appendix

In [130]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
plt.rc("font", size=14)
import scipy
import seaborn as sns
sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)
```

In [131]:
```
data = pd.read_csv("regdata.csv", header=0)
data= data.dropna()
print(data.shape)
print(list(data.columns))
```

```
(20, 5)
['score', 'Kate', 'prop', 'class avg', 'dev']
```

In [133]:
```
print(data)
```

```
    score      Kate    prop   class avg         dev
0       4    78.080  1.0000   88.267467  -10.187467
1       4   100.979  1.0000   54.414000   46.565000
2       4   161.271  1.0000   72.154267   89.116733
3       4   104.260  1.0000   62.805933   41.454067
4       4    39.683  0.9375   82.073429  -42.390429
5       3    76.783  1.0000   87.286333  -10.503333
6       3   133.847  1.0000   98.390467   35.456533
7       3   227.356  0.9375  114.743071  112.612929
8       3    60.124  0.9375   97.302786  -37.178786
9       4    75.645  0.8750  102.075308  -26.430308
10      4    54.998  1.0000  114.473733  -59.475733
11      4    72.990  0.7500   88.652182  -15.662182
12      4   124.786  0.9375   67.273643   57.512357
13      4   125.938  0.7500   81.152909   44.785091
14      4   155.042  0.8750   73.708385   81.333615
15      4   134.977  0.9375   89.804170   45.172830
16      3    60.861  0.8750   84.437000  -23.576000
17      3    70.051  1.0000   71.275867   -1.224867
18      4    67.695  0.7500   67.883643   -0.188643
19      3   199.372  0.3750  135.849800   63.522200
```

In [134]:
```
data['score'].unique()
```

Out[134]:
```
array([4, 3], dtype=int64)
```

In [135]:
```
data['score']=np.where(data['score']==4,1,data['score'])
data['score']=np.where(data['score']==3,0,data['score'])
```

In [136]:
```
data['score'].unique()
```

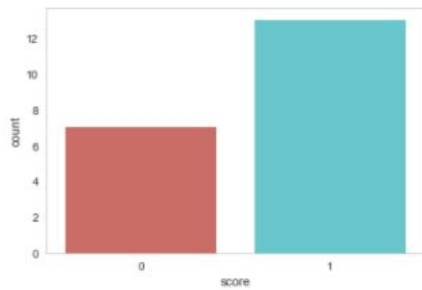
Out[136]:
```
array([1, 0], dtype=int64)
```

```
In [135]: data['score']=np.where(data['score']==4,1,data['score'])
          data['score']=np.where(data['score']==3,0,data['score'])
```

```
In [136]: data['score'].unique()
```

```
Out[136]: array([1, 0], dtype=int64)
```

```
In [155]: sns.set_style("whitegrid", {'axes.grid' : False})
          sns.countplot(x='score', data=data, palette='hls')
          plt.show()
```



More 4's than 3's - yay!

```
In [138]: data.groupby('score').mean()
```

Out[138]:

score	Kate	prop	class avg	dev
0	118.342000	0.875000	98.469332	19.872668
1	99.718769	0.908654	80.364544	19.354226

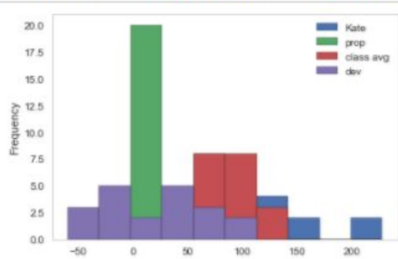
Observation: I've received 3's when I've spoken more More 4's when more students in class Spoke more on average for both conditions

```
In [139]: %matplotlib inline
```

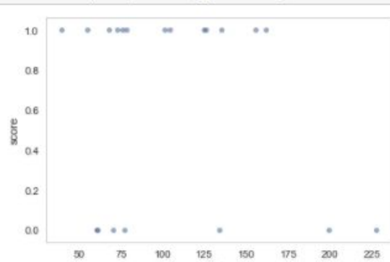
```
In [140]: data.isnull().sum() #no missing values
```

```
Out[140]: score        0
          Kate         0
          prop         0
          class avg    0
          dev          0
          dtype: int64
```

```
In [141]: ax = data.plot(x='score', y=None,kind='hist',edgecolor="k")
```



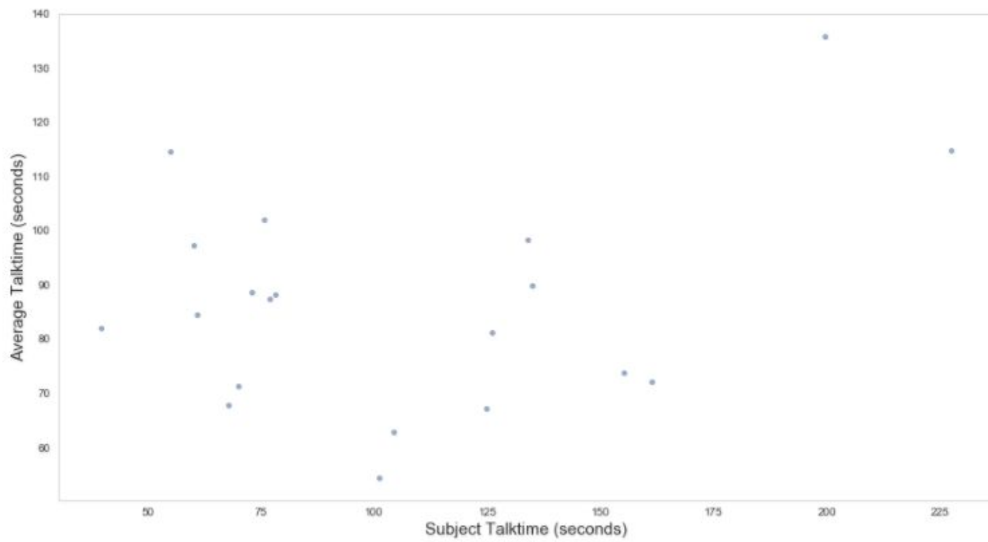

```
In [142]: ax0 = data.plot(x='Kate', y='score',kind='scatter',edgecolor="k", alpha=0.6)
```




```
In [163]: ax1 = data.plot(x='Kate', y='class avg',kind='scatter',edgecolor="k", alpha=0.6, figsize = (15,8))
          plt.ylabel('Average Talktime (seconds)', size=15)
          plt.xlabel('Subject Talktime (seconds)',size=15)
```

```
Out[163]: <matplotlib.text.Text at 0x1ddc5115978>
```

```
In [163]: ax1 = data.plot(x='Kate', y='class avg',kind='scatter',edgecolor="k", alpha=0.6, figsize = (15,8))
          plt.ylabel('Average Talktime (seconds)', size=15)
          plt.xlabel('Subject Talktime (seconds)',size=15)

Out[163]: <matplotlib.text.Text at 0x1ddc5115978>
```




```
In [119]: scipy.stats.pearsonr(data['Kate'],data['class avg'])

Out[119]: (0.30752558335181795, 0.18717682982606215)
```

The value of R is 0.3075. Although technically a positive correlation, the relationship between variables is weak (nb. the nearer the value is to zero, the weaker the relationship).

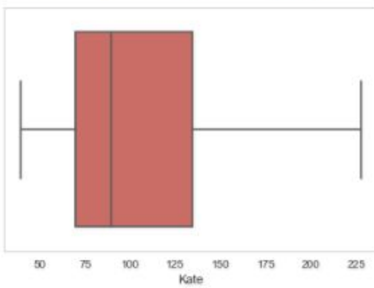
The value of R2, the coefficient of determination, is 0.0946.

```
In [154]: data.describe(percentiles=None, include=None, exclude=None)
```

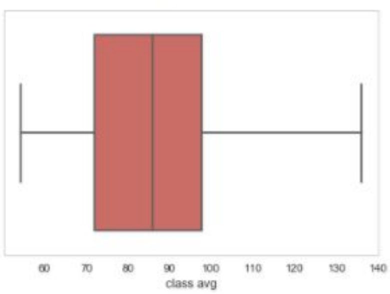
Out[154]:

	score	Kate	prop	class avg	dev
count	20.00000	20.000000	20.000000	20.000000	20.000000
mean	0.65000	106.236900	0.896875	86.701220	19.535680
std	0.48936	50.702133	0.150759	19.862080	48.433680
min	0.00000	39.683000	0.375000	54.414000	-59.475733
25%	0.00000	69.462000	0.875000	71.934667	-17.640636
50%	1.00000	89.529500	0.937500	85.861667	17.633945
75%	1.00000	134.129500	1.000000	97.574706	49.301839
max	1.00000	227.356000	1.000000	135.849800	112.612929

```
In [153]: sns.boxplot(x='Kate',y=None,data=data, palette='hls')
Out[153]: <matplotlib.axes._subplots.AxesSubplot at 0x1ddc5d9c7f0>
```



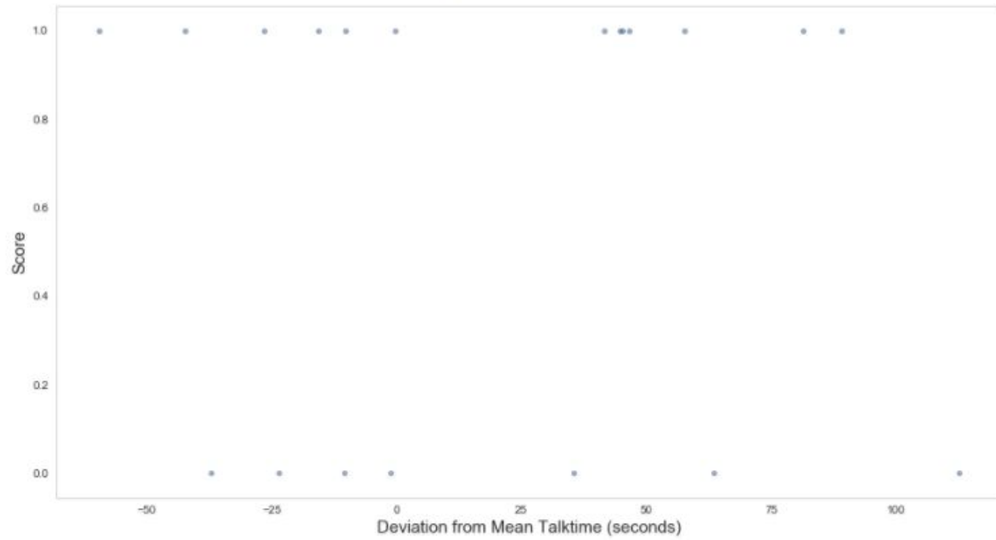

```
In [149]: sns.boxplot(x='class avg',y=None,data=data, palette='hls')
Out[149]: <matplotlib.axes._subplots.AxesSubplot at 0x1ddc5afc6a0>
```



```
In [88]: scipy.stats.spearmanr(data['Kate'],data['score'],axis=0)
Out[88]: SpearmanrResult(correlation=-0.045448911641126979, pvalue=0.84910230630735439)
```

The value of R is -0.04545 and the two-tailed value of P is 0.8491. By normal standards, the association between the two variables would not be considered statistically significant.

```
In [166]: ax1 = data.plot(x='dev', y='score',kind='scatter',edgecolor="k", alpha=0.6, figsize = (15,8))
          plt.ylabel('Score', size=15)
          plt.xlabel('Deviation from Mean Talktime (seconds)',size=15)


          scipy.stats.spearmanr(data['dev'],data['score'],axis=0)
Out[166]: SpearmanrResult(correlation=0.027269346984676195, pvalue=0.90914233534718769)
```



The value of R is 0.02727 and the two-tailed value of P is 0.90914. By normal standards, the association between the two variables would not be considered statistically significant.

```
In [102]: scipy.stats.pearsonr(data['Kate'],data['prop'])
Out[102]: (-0.31174953029631614, 0.18088039105040604)
```

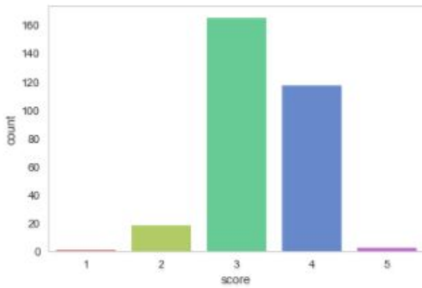
The value of R is -0.3117. Although technically a negative correlation, the relationship between your variables is only weak (nb. the nearer the value is to zero, the weaker the relationship).

The value of R2, the coefficient of determination, is 0.0972.

```
In [100]: scipy.stats.spearmanr(data['prop'],data['score'],axis=0)
Out[100]: SpearmanrResult(correlation=-0.056991083017872604, pvalue=0.81137368813261568)
```

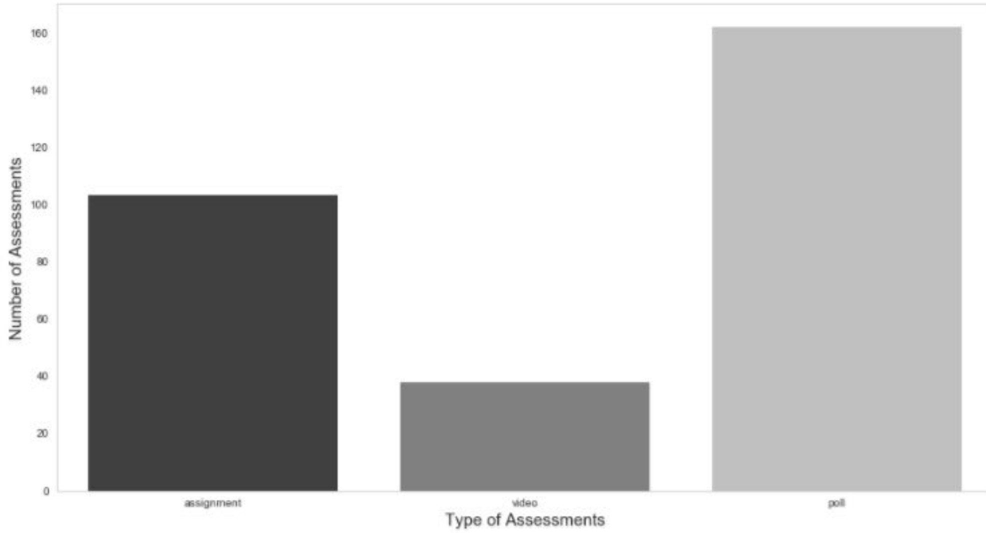
The value of R is -0.05699 and the two-tailed value of P is 0.81137. By normal standards, the association between the two variables would not be considered statistically significant.

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        plt.rc("font", size=14)
        import scipy
        import seaborn as sns
        sns.set(style="white")
        sns.set(style="whitegrid", color_codes=True)

In [2]: data = pd.read_csv("assessmentype.csv", header=0)
        data= data.dropna()
        print(data.shape)
        print(list(data.columns))

        (334, 2)
        ['type', 'score']

In [3]: data = data[data.score != 0] #removing HC scores with missing values

        data['score'].unique()

Out[3]: array([4, 3, 2, 5, 1], dtype=int64)

In [4]: sns.set_style("whitegrid", {'axes.grid' : False})
        sns.countplot(x='score', data=data, palette='hls')
        plt.show() #normal distribution
```




```
In [5]: fig, ax = plt.subplots(figsize=(15,8))


        sns.countplot(x='type', data=data, palette='gray', ax=ax)
        plt.ylabel('Number of Assessments', size=15)
        plt.xlabel('Type of Assessments',size=15)
        plt.show()
```

In [5]:
```
fig, ax = plt.subplots(figsize=(15,8))


sns.countplot(x='type', data=data, palette='gray', ax=ax)
plt.ylabel('Number of Assessments', size=15)
plt.xlabel('Type of Assessments',size=15)
plt.show()
```



In [6]:
```
data.groupby('type').mean()
```

Out[6]:

type	score
assignment	3.504854
poll	3.265432
video	3.157895

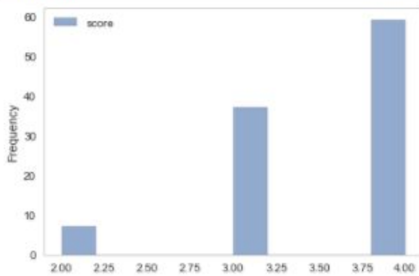
observation: higher hc scores in assignments, least scores in video

In [7]:
```
assign = data.loc[data['type'] == 'assignment']
assign.describe(percentiles=None, include=None, exclude=None)
```

Out[7]:

	score
count	103.000000
mean	3.504854
std	0.624245
min	2.000000
25%	3.000000
50%	4.000000
75%	4.000000
max	4.000000

In [8]:
```
assign.plot(x='type', y='score',kind='hist',edgecolor="k", alpha=0.6)
plt.show()
```

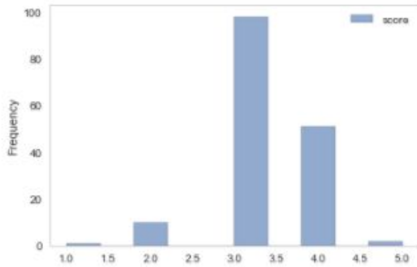


In [9]:
```
poll = data.loc[data['type'] == 'poll']
poll.describe(percentiles=None, include=None, exclude=None)
```

Out[9]:

	score
count	162.000000
mean	3.265432
std	0.618486
min	1.000000
25%	3.000000
50%	3.000000
75%	4.000000
max	5.000000

In [10]:
```
poll.plot(x='type', y='score',kind='hist',edgecolor="k", alpha=0.6)
plt.show()
```

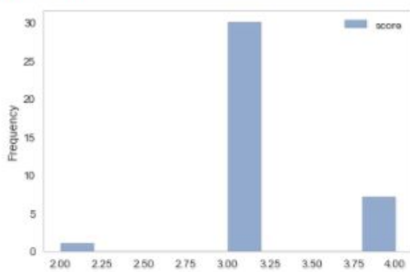


In [11]:
```
vid = data.loc[data['type'] == 'video']
vid.describe(percentiles=None, include=None, exclude=None)
```

Out[11]:

	score
count	38.000000
mean	3.157895
std	0.436591
min	2.000000
25%	3.000000
50%	3.000000
75%	3.000000
max	4.000000

In [12]:
```
vid.plot(x='type', y='score',kind='hist',edgecolor="k", alpha=0.6)
plt.show()
```



In [13]:
```
data2 = pd.read_csv("progress.csv", header=0)
data2= data.dropna()
print(data2.shape)
print(list(data2.columns))
```

```
(303, 2)
['type', 'score']
```

```
#include<bits/stdc++.h>
using namespace std;


int main() {
    freopen("C:\\Users\tanha\\AppData\\Local\\Packages\\Microsoft.SkypeApp_kzf8qxf38zg5c\\LocalState\Downloads\\in.txt","r",stdin);
    freopen("C:\\Users\tanha\\AppData\\Local\\Packages\\Microsoft.SkypeApp_kzf8qxf38zg5c\\LocalState\Downloads\\out.txt","w",stdout);
    string str;
    getline(cin,str);
    int doshok = str[8] - '0';
    int ekok = str[9] - '0';
    int previousNumber = doshok * 10 + ekok;
    int previousMarks = (float)(str[str.length()-1] - '0');
    float count = 1;
    float sum = previousMarks;
    string prevString = str;
    while(getline(cin,str)){
        int doshok = str[8] - '0';
        int ekok = str[9] - '0';
        float marks = (float)(str[str.length()-1] - '0');
        int currentNumber = doshok * 10 + ekok;
        if(currentNumber == previousNumber){
            sum+=marks;
            count++;
//            cout <<"COUNT: " <<  count << " SUM" << sum << "String: " << str << endl;
            prevString = str;
        }
        else{
            previousNumber = currentNumber;
//            printf("ADDED: %s PREV: %d SUM: %0.2f COUNT:%f Str: %s \n",str.c_str(),previousNumber,sum/count,count,prevString.c_str());
//            printf("%s\n",prevString.substr(0,10).c_str());
            printf("%0.2f\n",sum/count);

            prevString = str;
            count = 1;
            sum = marks;
        }




//        printf("%d %0.2f\n",currentNumber,marks);
//
//        printf("%s\n",str.c_str());
    }
    printf("%s %0.2f\n",prevString.substr(0,10).c_str(), sum/count);
    return 0;
}
```