TODO:

; Move all figures and tables to the end.

; Go through the comments. Some indicate rewording. At least one discusses a new experiment.

; Consider removing Amy’s rating, instead using just a scatter plot of each student’s rate of completion vs exam grade. Discretizing decreases the transparency as shown by the more complicated methods (Freeman-Halton extension of Fisher’s exact test vs Spearman correlation, for example)

Title: Do reflective students learn more in the Emergency Department?

Abstract:

**Background:** Medical schools have begun to incorporate self-reflection exercises into their curricula. It is thought that these exercises lead to deeper mastery of the material and better academic performance. There are few data supporting this hypothesis.

**Objectives:** We evaluated the relationship between the degree of reflection after a student’s shift in an emergency room and that student’s final grade.

**Methods:** We conducted a retrospective case series by analyzing the performance and reflective statements of 116 students who participated in an emergency medicine clerkship at two clinical sites from 2013-14. After each shift, an attending emergency medicine physician evaluated the student and the student could complete an optional reflection freetext section. We analyzed the correlation between the final grade, expressed in quartiles, and the degree to which the student completed the reflection using the Freeman-Halton extension of Fisher’s exact test. We used latent Dirichlet allocation to identify which topics each quartile discussed. We quantified the similarity between those topics using Jaccard similarity. For all statistics except the Freeman-Halton D, statistical significance was determined using bootstrapping.

**Results:** Of the 145 possible records, 116 were included for analysis. The other 29 were excluded as they were visiting students. Two EM physicians graded the rate of completion of the self-reflection, demonstrating moderate agreement in their assessment (Cohen’s kappa = 0.55). The assessments of both raters were significantly correlated with final grade. (p=0.006 and p=0.008.) Students with the lowest grades wrote the least amount of reflection. (Metrics)

**Conclusions:** The degree and quality of student reflection significantly correlate with final grade in an emergency clerkship designed for fourth year medical students. In future, as faculty preform the evaluations, they can encourage more insightful reflection from the students to improve their performance in the clerkship.

**Introduction**

Undergraduate medical education is becoming, increasingly, self-directed (CITATION). Reflection is an essential aspect of self-regulated learning, and may also be required to develop a therapeutic relationship and professional expertise**[[1]](#footnote-1)**.

- Importance of self-reflection in self-regulated learning

- Increasing prevalence of self-regulated learning in medical school

- Methodologic difficulty of assessing reflection, especially in retrospective surveys; we captured the thoughts immediately after each shift.

**Methods**

We conducted a retrospective analysis of all student performances from October 2014 to October 2015. Figure 1 summarizes our study design.

**Grading.** At the end of each shift, an attending physician in Emergency Medicine evaluated the student according to the scheme in Table 1. Medical students rotate in one of four hospitals, one city hospital, two community hospitals, and one academic hospital. We only included fourth-year students from our institution taking the Emergency Medicine clerkship for the first time. We did not include visiting students.

**Software.** All analysis was written by MC and performed in Python**[[2]](#footnote-2)**. Natural language processing was performed using the Natural Language Toolkit for Python, version 3.0**[[3]](#footnote-3)**. Bootstrapping and the calculation of Jaccard similarity were performed using NumPy/SciPy**[[4]](#footnote-4)**. Figures were made using the matplotlib**[[5]](#footnote-5)** plugin. Multinomial Naive Bayes Classification was performed using NLTK and sci-kitlearn**[[6]](#footnote-6)**. All code used to analyze and generate the figures as well as supporting documents are available at the following GitHub repository: [**https://github.com/mac389/leuthauser**](https://github.com/mac389/leuthauser).

**Data Acquisition.** After each rotation a student fills out an evaluation form, which is a page in a booklet. At the end of the rotation, every student returned his booklet to the directors of the Emergency Medicine Clerkship (AL, BH). Author MC transcribed the booklets into a database. In transcribing, all medical abbreviations and contractions were replaced with their long form and all illegible comments were ignored. A deidentified version of the database is available at our GitHub repository. The text of each student comment was processed as follows:

1. All text converted to lower case
2. Comments tokenized into words
3. Stopwords removed
4. Remaining words lemmatized

*Stopwords.* The term *stopwords* refers to words that occur frequently in a corpus but are unlikely to be informative. The removal of stopwords is a common preprocessing step in natural language processing. It can increase the sensitivity and specificity of analyses**[[7]](#footnote-7)**. The list of stopwords depends on the task. The list of stopwords we used is available at our repository. It is an amalgamation of the English stopwords list in the Natural Language Tooklit 3.0 package for Python and the 10,000 most frequently occurring words in the transcript of all episodes of *The Simpsons*.

*Lemmatization.* The term *lemmatization* refers to the mapping of all inflected forms of a word to a base form so that they can be analyzed as a single item. For example lemmatization maps “infect”, “infected”, “infection”, “infections”, and so on to “infect”. To lemmatize words in our study we used the *WordNetLemmatize* function in NLTK 3.0. This function is a thin wrapper to WordNet’s *morphy* function, which removes all suffixes that occur in the WordNet database. *WordNetLemmatize* is more accurate if it known the part of the speech of the word it is asked to lemmatize. For example, *patient* and *patients* should only be considered one item if *patient* is a noun. To identify the part of speech of each word, we used the *pos\_tag* function in NLTK 3.0. *Pos\_tag* is trained on the treebank corpus*[[8]](#footnote-8)*.

*Tokenization.* The term *tokenization* refers to breaking a string of words into those words. Tokenization can be difficult when abbreviations and nonstandard punctuation are used. We used the *word\_tokenize* function in NLTK 3.0. This tokenizer uses regular expressions and is appropriate for pieces of text that do not have emoticons nor use contractions extensively.

*Jaccard similarity.*The Jaccard similarity**[[9]](#footnote-9)** quantifies the similarity between two sets of objects. The Euclidean distance, in contrast, quantifies the distance between an ordered series of numbers. The Jaccard similarity is defined as the ratio of number of objects two sets have in common to the total number of unique objects across both sets.

*Bootstrapping.* In statistics *boostrapping* refers to a process of resampling without replacement to generate an empirical probability density function**[[10]](#footnote-10)**. It allows the estimation of the statistical significance of a parameter when the underlying distribution is not known. In this paper we use it to estimate the statistical significance of Jaccard similarities.

**Results**

*Demographics.* Figure 1 shows the distributions of faculty ratings (left) and exam performances (right). Hyphenated categories, such as *Reporter-Interpreter* refer to evaluations where the attending circled two adjacent categories evenly. Figure 2 shows the fifty most common words and nine most common bigrams, two-word phrases, in all student comments that remained after the processing described in the **Methods** section. As an example, after processing the phrase *loss of consciousness* becomes (*loss*, *consciousness*). That phrase has two unigrams *loss* and *consciousness* and one bigram *loss consciousness*.

*Correlation between reflection and exam performance.* Figure 3 shows the correlation between the fraction of comments each student completed (left) or the average lengths of each student’s comments and that student’s performance on the final exam. We considered a comment completed if the comment had at least one legible word. The correlation between the fraction of comments each student completed and exam performance was significant (two-tailed t-test; p=0.03). The correlation between the average length of comments and exam performance was not significant (two-tailed t-test; p=0.16). There are two clusters in the right panel of Figure 3. The cluster of hollow circles corresponds to students who commented on their experience more than half of the time. The cluster of solid circles corresponds to those who commented less than half of the time. The median test scores of the completes and non-completers, 83 +/-4 and 79 +/- 3 (median +/- interquartile range) are not significantly different. The 95% confidence intervals for the medians overlap (Figure 5). The clusters do come from different distributions (Kolmogorov-Smirnov test; D statistic 0.5; p=.005). With a larger sample the difference may become statistically significant.

*Analysis of text.* Figure 3 shows the most common words in the clusters Figure 2 identified. The Jaccard similarity between these two clusters was 0.06. Figure 4 shows the most common words in each category described in Table 1. We excluded student rated as inadequate because too few students were rated as “Inadequate”. We also excluded hybrid categories.

Table 2 shows the Jaccard similarity between all pairs of panels in Figure 4. It suggests that those rated “Manager” used more words in common with those rated “Superior” than did those rater “Reporter” or “Interpreter”.

**Conclusions**

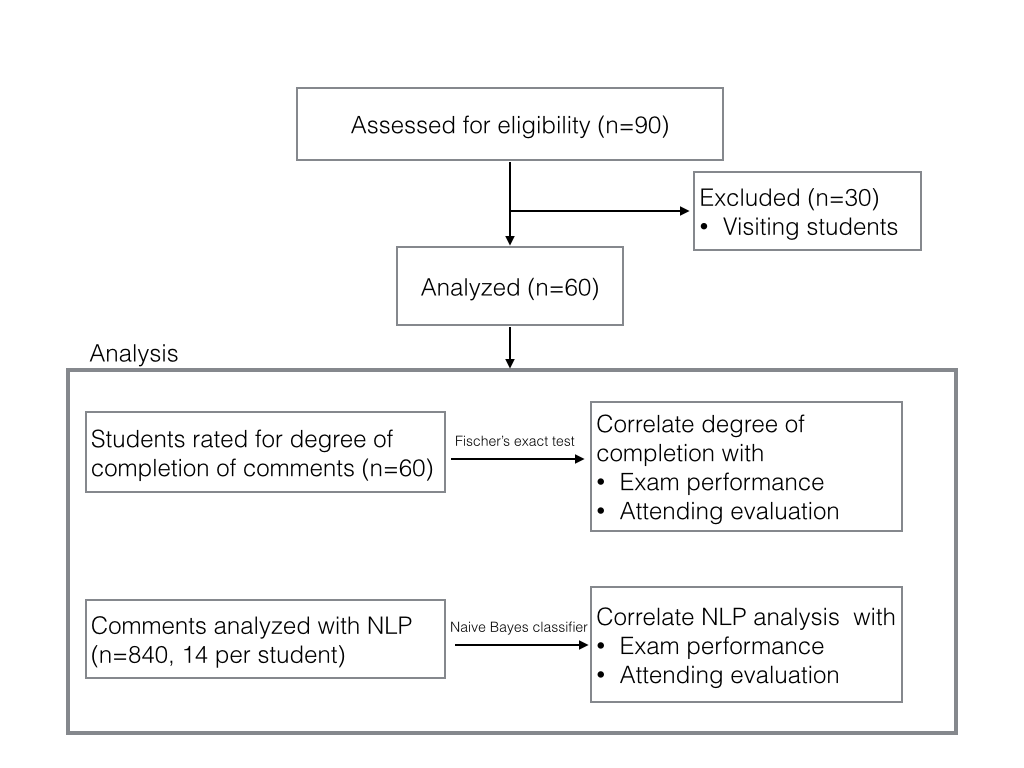
The main result of this study is that the number of reflective comments a medical student makes after his shift correlates with his performance on a final written exam in an Emergency Medicine clerkship. Students who wrote more reflective comments scored higher on the final exam, although this increase was not statistically significant. Counting comments with single legible words as “complete” may have diluted our statistical power. The association between completion of a reflection exercise and exam performance may reflect an underlying attribute, such as diligence.

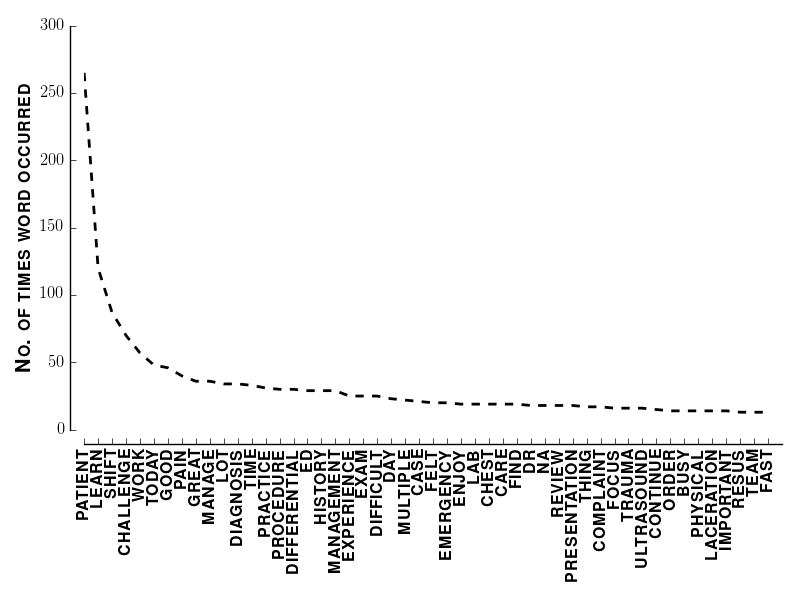
A secondary result is that “Reporters”, “Interpreters”, “Managers”, and “Educators” use different words to describe their reflections**[[11]](#footnote-11)**. The word frequencies were tabulated once the attending evaluations were known. We could not determine the statistical significance of this association. A chi-square test is inappropriate. The observations are not independent. Each student generated 14 comments and attending evaluations. It is not known, however, whether the students wrote their reflections after each shift or in bunches. Over 200 were evaluated as “superior”. Disproportionate representation of some categories at the expense of others makes it harder to find distinguishing features of each category. We excluded hybrid categories. In many cases it was unclear whether the attending meant to circle both categories, indicate performance between two categories, or meant to circle only one category. Excluding those comments may have decreased the power. There were insufficient data to train a naive Bayes classifier to predict attending evaluations from student comments.

We could not control for seasonality. We had only one year of data. We did not have data on the specialties students were interested in. Medical students looking to pursue residency in emergency medicine may perform better than students interested in other specialties. They may schedule the clerkship earlier than other students to obtain letters of recommendation, or because, at our institutions, students must rotate through our emergency department before they rotate at other institutions.

This study is the first to use natural language processing to quantify the relationships between how medical students perceive their performance, how attendings perceive their performance, and student performance on an objective measure, a final exam. Natural language processing has been used previously in more restrictive cases, such as to partially automate detecting when medical students have met procedural requirements**[[12]](#footnote-12)**. This study suggests a large role for natural language processing in medical education.

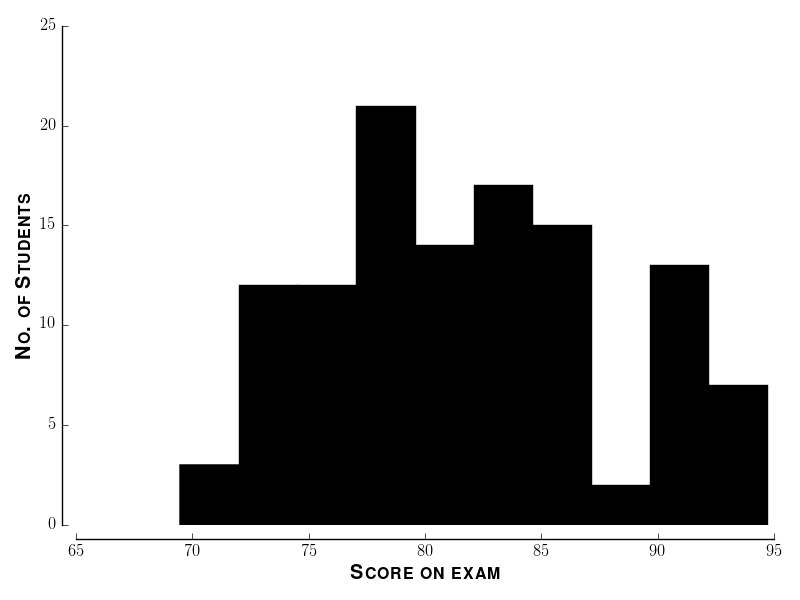
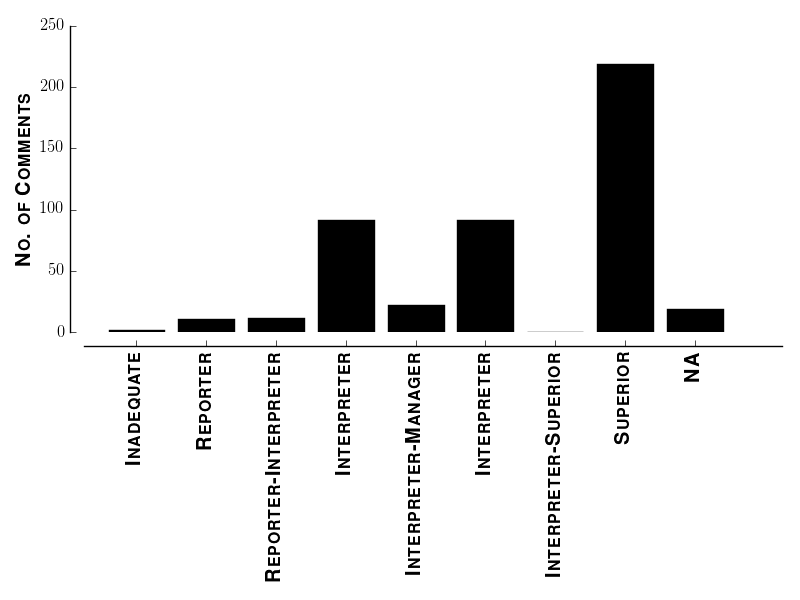
**Figure 1. Study design.**



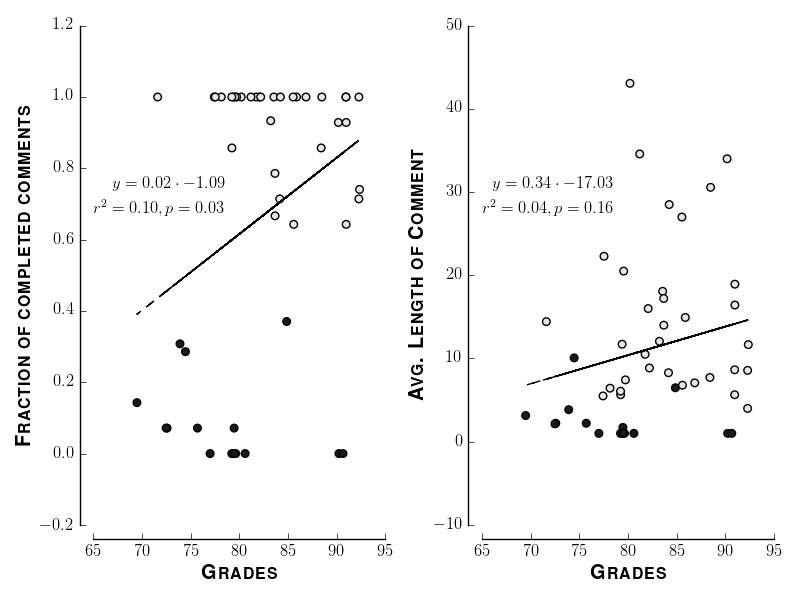


| Bigram (frequency) |
| --- |
| learn lot (23) |
| differential diagnosis (19) |
| chest pain (14) |
| manage patient (12) |
| multiple patient (11) |
| patient learn (9) |
| emergency department (8) |
| incision drainage (8) |
| abdominal pain (8) |

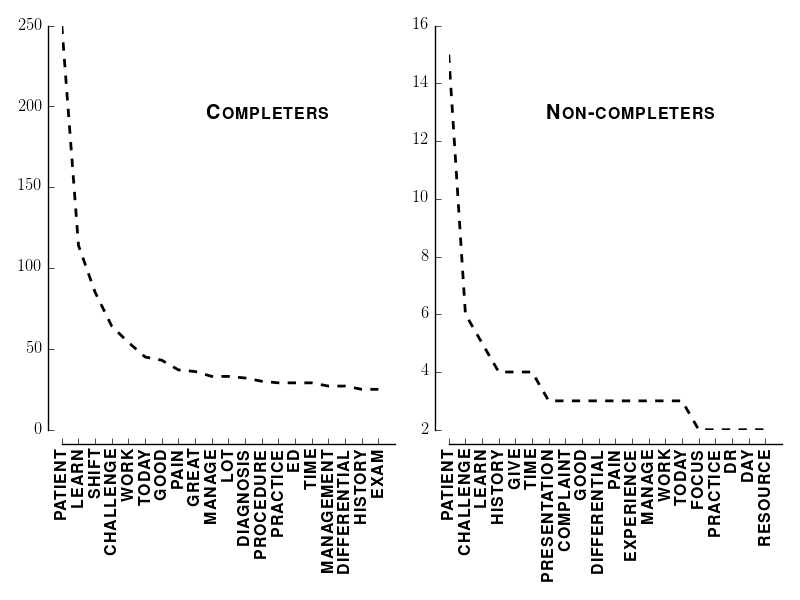
**Figure 1. Left:** Most common words in all student comments. **Right:** Most common bigrams.



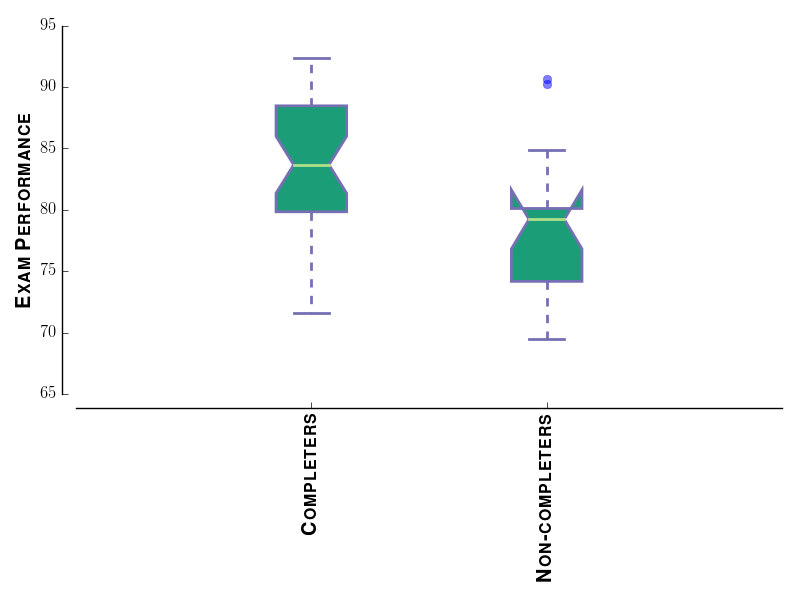
**Figure 2. Study demographics. Left:** Distribution of attending ratings. Hyphenated ratings indicate that an attending circled two categories. **Right:** Distribution of final exam grades.

****

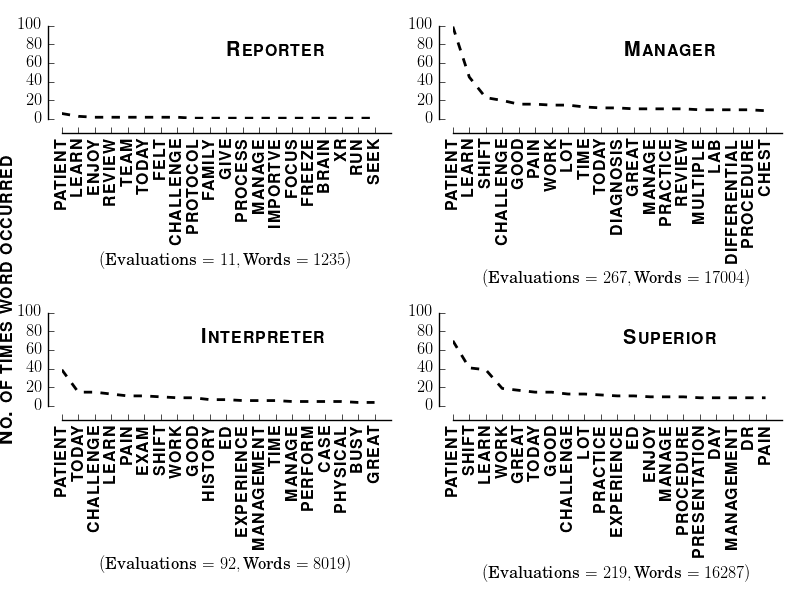
**Figure 3. Correlation between exam performance and reflection.** **Left:** Scatter plot of grades versus fraction of completed comments. Each point represents one student. Dashed line indicates regression of fraction of completed reflections against grades. **Inset:** Top line shows equation of regression line. Bottom line shows coefficient of determination and p-value that the slope of the regression line is significantly different from zero. **Right:** Scatter plot of grades versus average length of comments. Each point represents one student. Dashed line and inset indicate the same as in the left panel. In both panels solid circles represent those who completed less than half of the comments. Hollow circles represent those who completed more than half of the comments.



**Figure 4. Word frequencies. Left:** Twenty most common words used by those who completed more than half of their comments. **Right:** Twenty most common words used by those who completed less than half of their comments.



**Figure 5.**

**Table 2.** Jaccard similarity.

**Table 1. RIME Rating scheme.**

|  |  |
| --- | --- |
| Category | Description |
| Inadequate | Well below expected |
| Reporter | Obtains and reports information, beginning to interpret |
| Interpreter | Consistently access data. Actively participates in care |
| Manager | Suggests treatment options based on data. Deep knowledge of his/her patients |
| Superior | Fewer than 5% of students. Works at or above the level of EM intern. Cana manage patients independently |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Reporter | Interpreter | Manager | Educator |
| Reporter |  |  |  |  |
| Interpreter | 0.989 (p=0.436) |  |  |  |
| Manager | 0.1494 (p=0.2694) | **0.3699 (p<0.001)** |  |  |
| Educator | 0.1364 (p=0.3166) | **0.3158 (p<0.01)** | **0.4286 (p<0.01)** |  |

1. Sandars, J. (2009). The use of reflection in medical education: AMEE Guide No. 44. Medical teacher, 31(8), 685-695. [↑](#footnote-ref-1)
2. Van Rossum, G., & Drake, F. L. (2003). Python language reference manual (p. 144). Network Theory. [↑](#footnote-ref-2)
3. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. " O'Reilly Media, Inc.". [↑](#footnote-ref-3)
4. Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering, 13(2), 22-30. [↑](#footnote-ref-4)
5. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science and engineering, 9(3), 90-95.

   Chicago [↑](#footnote-ref-5)
6. Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." The Journal of Machine Learning Research 12 (2011): 2825-2830. [↑](#footnote-ref-6)
7. Rajaraman, A.; Ullman, J. D. (2011). "Data Mining". Mining of Massive Datasets. pp. 1–17. doi:10.1017/CBO9781139058452.002. ISBN 9781139058452. edit [↑](#footnote-ref-7)
8. Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 19(2), 313-330. [↑](#footnote-ref-8)
9. Lapointe, F. J., & Legendre, P. (1994). A classification of pure malt Scotch whiskies. Applied Statistics, 237-257. [↑](#footnote-ref-9)
10. Davison, A. C.; Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. [↑](#footnote-ref-10)
11. A new vocabulary and other innovations for improving descriptive in-training evaluations. Acad Med. 1999 Nov;74(11):1203-7. [↑](#footnote-ref-11)
12. Denny, J. C., Bastarache, L., Sastre, E. A., & Spickard, A. (2009). Tracking medical students’ clinical experiences using natural language processing. Journal of biomedical informatics, 42(5), 781-789. [↑](#footnote-ref-12)