Research Proposal for ED6460 Summer 2017 Sandy Davis

Introduction:

In the United States, and in most countries around the world, females and certain racial minorities are significantly under-represented in STEM university programs and STEM professional fields (Huang & Walter, 2000; NSF, 2013; McPhee, et al, 2013; "Women in Science," 2016). Recruitment efforts at the university level have shown some promise, however retention of females and minorities continues to be a challenge in many programs (Koenig, 2009, Blickenstaff, 2005). More insight is needed into why certain demographic groups are less likely to complete STEM university programs.

My project is centered around two research-based assumptions. The first assumption is that social learning and a feeling of connection are keys to success in academic programs (Wright, et al 2014; McKay 2017; Foltz, et al 2014; May & Chubin, 2003). The second assumption is that certain groups experience feelings of inadequacy that can negatively impact their academic performance and participation (Steele, 2010; Spencer, et al, 1999). Lower levels of personal confidence have been found specifically in females in higher education STEM programs (Chachra and Kilgore, 2009; Kamas, et. al. 1993; Fisher, et al 1997; Hayes, 2017).

In the OMSCS course the primary means of community and communication is each course's Piazza forum. Based on research around Imposter Syndrome and stereotype threat it is plausible that certain demographic groups are posting to Piazza at a lower than expected rate because of a lack of confidence and perceived feelings of not belonging in a top graduate CS program. If that is the case, then these groups' social learning opportunities are being limited and their academic outcomes may be lower than what they could be with full participation.

My hypothesis is that certain sub-groups in the OMSCS population are under-participating in the Piazza forum. These sub-groups are females, students who did not major in CS, students who are not currently working in a Computer Science related field, students who have relatively few years of programming experience, and students who are newer to the OMSCS program (having completed 3 or fewer courses).

I will be analyzing actual Piazza forum data for an OMSCS class, as well as analyzing survey response data related to self- reported Piazza usage and confidence.

Research Question and Approaches

My guiding research question is: "Do certain sub-groups within the OMSCS student population under-participate on the Piazza forum? If so, what are the reasons for their lack of participation?". I will use two investigation approaches to try to answer this question. The first approach is to analyze actual Piazza forum data from a recent OMSCS course. The second approach is to collect and analyze more qualitative data via a survey sent to OMSCS students.

Approach 1:

Did a user's gender or current occupation significantly impact his or her Piazza participation in a recent OMSCS course based on actual forum data analysis?

Data source:

Piazza forum data from a recent OMSCS course.

Independent variables:

- Student's gender
- Student's occupation

Dependent variables:

- Number of original notes created by the student
- Number of original questions created by the student
- Number of original polls created by the student

- Number of replies/follow-ups created by the student
- Average number of words per post for the student

Proposed Methodology:

I will use Piazza forum data from Knowledge-Based Artificial Intelligence Spring 2017. Raw post data will be accessed via an unofficial Piazza API and includes the metadata associated with each original and follow-up post. Each post can be downloaded as a json file. Once all unique IDs have been collected I will use my front-end Piazza access for the course to view 1) actual names and photos of each student and 2) the Introductions posts for the course.

Gender will be determined using a combination of student name and photo associated with the ID. If the name is gender-specific and/or the photo clearly implies a gender then a gender code of male or female will be assigned. In cases where the name is associated with both genders and the photo does not imply a gender then a gender code of 'undetermined' will be assigned.

Occupation will be determined using the 'Introductions' posts that were created at the start of the class. In those posts a majority (184 out of 197) of the users identified themselves to the class. The introductory question prompt included career, so most replies include mention of occupation. If the respondent indicates a career directly related to Computer Science (software developer, software architect, software engineer, web developer, data scientist, db administrator or designer, computer security analyst) then an occupation code of "CS related" will be assigned. If the respondent indicates an occupation other than those in the CS related category then an occupation code of "non-CS related" will be assigned. If the type of work is not clear or not mentioned then an occupation code of "unknown/not specified" will be assigned.

I will use Python to import the Piazza post data files and determine posting activity by user ID. This data will be exported to a csv/Excel file for full analysis using a tool such as R-Studio.

The first layer of analysis will be simple comparison of means and standard deviations across the sample. I will compare the known percentage of females in the course with the percentage of posts by female students, then the known percentage of males in the course with the percentage of posts by male students. I will do the same type of comparison for the occupation category. This analysis will be descriptive only and not prescriptive. In order to determine statistical significance paired data will be used to run additional tests. If the data is found to be normal then a basic t-test will be run against the data to compare the mean number of posts (by type) across gender and occupation. If the data is not found to be normal then a Mann-Whitney U test will be used.

There is a chance that gender is not associated with posting activity when occupation is controlled for, or vice versa. In order to ensure that the independent variable in question is actually driving the results I will use hierarchical regression. This part of the research will support the internal validity of the findings.

If my data shows a significant difference in the mean number of posts by females and mean number of posts by males, or a significant different in the mean number of posts by those with a CS background vs a non-CS background, then the findings are generalizable and have predictive power. This means they will have external validity. If there is no statistical significance then the data may still show some interesting descriptive findings but cannot be generalized across the entire OMSCS population.

Approach 2:

How are factors such as gender, age, English language fluency, occupation, years of programming experience, and number of classes completed in the OMSCS program related to self-reported Piazza involvement, experience of disrespect on the Piazza forum, and sense of belonging in the overall OMSCS community?

Data source:

Research Proposal for ED6460 Summer 2017 Sandy Davis

Survey responses from OMSCS Google Plus community and OMSCS Nerdy Bones Google Plus community.

Independent variables:

- Gender
- Age
- Occupation
- English language fluency
- Years of programming experience
- Number of OMSCS courses completed
- Use of collaboration tools other than Piazza

Dependent variables:

- Self-reported posting activity of forum notes, questions, polls
- Self-reported posting activity of forum replies/follow-ups
- Self-reported confidence related to posting on the forum
- Whether or not the student has felt disrespected or offended by another student on the forum
- Self-reported sense of belonging in the overall OMSCS community

Since the data results will be paired I will run t-tests (assuming normal data) to determine statistical significance and external validity between each of the independent variables and each of the dependent variables. If the data is not normal I will use a Mann-Whitney U test. Several independent variables will have more than two categories. In these cases I will use an ANOVA test instead of a t-test.

Once again to promote internal validity I will use hierarchical regression on the survey data. The survey question regarding collaboration tools other than Piazza was specifically added to support internal validity. Some students' lack of Piazza participation may have little to do with confidence and simply be due to their choice to use Slack instead of Piazza for collaboration.

Analyzing the findings:

My hope in using the two different data sources (forum data and survey data) is to see whether any patterns or associations in the forum data can be better explained with the survey data. It is one thing to find a significant relationship between two variables and another thing entirely to understand why that relationship exists. My hope is that the survey data explains some of the reasons behind any phenomena that the forum data illuminates.

Limitations:

A limitation of my proposal is the use of name and photo to determine the gender of each Spring 2017 KBAI user. Ideally gender would be self-reported however I do not have access to this data. In order to mitigate the risk of incorrectly coding a user's gender I will use both name and photo, and any case with any ambiguity will be coded as 'undetermined'.

Another limitation is time. With the condensed summer schedule there will be less time for survey response collection, data analysis, and write up. I believe I have created a realistic schedule, however timing will be tight. In order to mitigate the time risk I have already downloaded all data files needed and prototyped the Python code needed to process the data.

Task List and Timeline:

Task	Status	Estimated Completion Date	Comments
Conduct background research related to social learning theory, confidence impacts on performance, and existing research related to	Complete	11-Jun	Assignments 1-5, and personal question research
Finalize research question	Complete	11-Jun	Research Question: "Do certain sub-groups within the OMSCS student population under-participate on the Piazza forum? If so, what are some possible reasons for their lack of participation"
Explore options for actual Piazza forum research data.	Complete	6-Jun	The only scrubbed set of Piazza data is from KBAI Summer 2015. This data has already been analyzed against gender in Sims (2016) research. The only other option is to use data from a class I was enrolled in, therefore I chose Spring 2017 KBAI. Through mentor I learned about unofficial API that allows access to raw post data.
Use Piazza API to pull down json files containing Piazza KBAI Spring 2017 semester data.	Done	6-Jun	Used https://github.com/hfaran/piazza-api for API calls
Convert json files to csv	Done	10-Jun	
Write method to process csv files and collect post activity data for each unique user	Done	10-Jun	
Match each unique user in KBAI Spring 2017 with gender and instructor codes	Done	12-Jun	
Match each unique user in KBAI Spring 2017 with occupation category	Done	12-Jun	
Run data collection method and finalize data file	Not started	2-Jul	DELIVERABLE FOR MILESTONE 1
Use mean comparison to generate descriptive statistics for gender and occupation.	Not started	2-Jul	DELIVERABLE FOR MILESTONE 1
Run t-test (or Mann-Whitney-U test if data not normal) to determine if there is statistically significant differences between posting activity of by gender, and by occupation category	Not started	2-Jul	DELIVERABLE FOR MILESTONE 1
Run hierarchical regression to determine internal validity.	Not started	2-Jul	DELIVERABLE FOR MILESTONE 1
Receive final approval for survey and proposal	In progress	20-Jun	
Send out survey	Not started	26-Jun	Deadline for responses July 7
Analysis of survey data using paired t-test (or Mann-Whitney U test if data not normal)	Not started	16-Jul	DELIVERABLE FOR MILESTONE 2
Run hierarchical regression to determine internal validity.	Not started	16-Jul	DELIVERABLE FOR MILESTONE 2
Create final report	Not started	30-Jul	
Create final presentation	Not started	30-Jul	

Task Phase Key
Gather Information
Collect and Analyze Piazza Forum Data
Collect and Analyze Survey Data
Process and Package Results

Research Proposal for ED6460 Summer 2017 Sandy Davis

Milestone 1:

For this deliverable I will create a ppt presentation showing the following:

- Process used to collect data from Piazza
- Process used to analyze data
- Descriptive statistical findings related to Piazza sampling
    - Mean
    - Standard Deviation
    - Skewedness
- Predictive statistical findings related to Piazza sampling
    - Paired t-test or Mann-Whitney U test
    - Hierarchical regression
- Discussion of results so far and next steps

Milestone 2:

For this deliverable I will create a ppt presentation showing the following:

- Process used to collect survey data
- Process used to analyze survey data
- Descriptive statistical findings related to survey results
    - Mean
    - Standard Deviation
    - Skewedness
- Predictive statistical findings related to survey results
    - Paired t-test or Mann-Whitney U test
    - ANOVA
    - Hierarchical regression
- Discussion of results so far and next steps

Goals of both milestones: Peer review of methods, peer insight into data findings, peer feedback of next steps.

Draft survey questions:

1. What is your age?
    a. Younger than 35
    b. 35 or older
2. What was your undergraduate major?
    a. Computer Science
    b. Engineering
    c. Other
3. Do you currently work as a software developer, software architect, software engineer, web developer, data scientist, database administrator or designer, or computer security analyst?
    a. Yes
    b. No
4. How would you rate your English language proficiency?
    a. Native speaker
    b. Non-native speaker, fluent
    c. Non-native speaker, proficient
5. What is your gender?
    a. Female

   b. Male
   c. Prefer not to answer
6. How many courses have you completed so far in the OMSCS program?
   a. 0 courses (newly admitted student or enrolled in first course)
   b. 1-3 courses
   c. 4-6 courses
   d. 7 or more courses
7. About how many years of programming experience do you have?
   a. 0 to 3 years of experience
   b. More than 3 years to 5 years of experience
   c. More than 5 years to 10 years of experience
   d. More than 10 years of experience
8. In your most recently completed OMSCS class about how many times did you **post a question, note, or poll** to the Piazza forum? If you completed two or more courses last semester please choose one course to base your answer on.
   a. Never
   b. 1-3 times
   c. 4-10 times
   d. More than 10 times
9. In your most recently completed OMSCS class about how many times did you **reply to another student's question or note** on the Piazza forum? If you completed two or more courses last semester please choose one course to base your answer on.
   a. Never
   b. 1-3 times
   c. 4-10 times
   d. More than 10 times
10. Besides Piazza, are there other ways you connect with students in your courses? (select all that apply)
   a. Slack or other online team communication platform (ex: HipChat, Rocket Chat)
   b. Google Plus community
   c. Email
   d. In-person study group
   e. I do not connect with other students in my courses outside of Piazza
11. On a scale from 1 to 4, how would you rate your confidence when it comes to posting questions, notes, polls, or replies to the Piazza forum?
   1 – not at all confident
   2 – slightly confident
   3 – moderately confident
   4 – extremely confident

Additional comments (optional):

12. Have you ever felt personally disrespected or offended by another student's post or reply on Piazza?
   a. No
   b. Yes

Additional comments (optional):

13. On a scale from 1 to 5 please rate the following statement: "I belong in the OMSCS program."
   1 – strongly disagree
   2 - slightly disagree
   3 – neither agree nor disagree
   4 – slightly agree

5 – strongly agree

Additional comments (optional):

Sources:

Clark Blickenstaff*, J. (2005). Women and science careers: leaky pipeline or gender filter?. *Gender and education*, *17*(4), 369-386.

Chachra, D. & Kilgore, D. (2009). Exploring Gender and Self-confidence in Engineering Students: A Multi-method Approach. Proceedings of the 2009 American Society for Engineering Education Conference.

Fisher, A., Margolis, J, & Miller, F. (1997). Undergraduate Women in Computer Science: Experience, Motivation and Culture. SIGCSE Bulletin, pp. 106–110.

Foltz, L. G., Gannon, S., & Kirschmann, S. L. (2014). Factors that contribute to the persistence of minority students in STEM Fields. *Planning for Higher Education*, *42*(4), 46.

Hayes, G. (2017). Educational Technology 6460 Project. Retrieved from https://files.t-square.gatech.edu/access/content/group/gtc-d1ef-61da-5b1b-aefd-963e8cb0a8bb/Past%20Semesters%27%20Projects/Spring%202017/Hayes%2C%20Genevieve/CS6460_ProjectPaper_GHayes.pdf

Huang, G., Taddese, N., & Walter, E. (2000). *Entry and persistence of Women and Minorities in College science and engineering education* (NCES Rep. No. 2000-601). Washington, DC: U.S. Government Printing Office.

Kamas, L., Paxson, C., Wang, A., & Blau, R. (1993). Ph.D. Student Attrition in the EECS Department at the University of California, Berkeley. University of California, Berkeley EECS Department.

Koenig, R. (2009). Minority retention rates in science are sore spot for most universities. *Science*, *324*(5933), 1386-1387.

MacPhee, D., Farro, S., & Canetto, S. S. (2013). Academic self-efficacy and performance of underrepresented STEM majors: Gender, ethnic, and social class patterns. *Analyses of Social Issues and Public Policy*, *13*(1), 347-369.

May, G. S., & Chubin, D. E. (2003). A retrospective on undergraduate engineering success for underrepresented minority students. *Journal of Engineering Education*, *92*(1), 27-39.

McKay, T. (May 24, 2017). Learning analytics: Harnessing data science to transform education. National Science Foundation presentation, University of Michigan.

National Science Foundation, National Center for Science and Engineering Statistics. (2013). Women, minorities, and persons with disabilities in science and engineering: 2013. Special report NSF 13–304. Arlington, VA: author. Retrieved from www.nsf.gov/statistics/wmpd/2013/pdf/nsf13304_digest.pdf

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, *35*(1), 4-28.

Steele, Claude. (©2010) *Whistling Vivaldi :and other clues to how stereotypes affect us* New York : W.W. Norton & Company.

Women In Science, Technology, Engineering, And Mathematics (STEM). (December 9, 2016). Retrieved from http://www.catalyst.org/knowledge/women-science-technology-engineering-and-mathematics-stem#footnote6_aq66mdu

Wright, M., McKay, T., Hershock, C., Miller, K., Tritz, J. (2014) Better Than Expected; Using Learning Analytics to Promote Student Success in Gateway Science. Change: The Magazine of Higher Learning 46.1: 28-34