Modeling Determinants of Undergraduate Computing Students' Participation in Internships

Megan Wolf
Computer & Info. Science & Engineering
University of Florida
me@meganwolf.com

Amanpreet Kapoor
Engineering Education
University of Florida
kapooramanpreet@ufl.edu

Charlie Hobson
Computer & Info. Science & Engineering
University of Florida
cdhobson@proton.me

Christina Gardner-McCune
Computer & Info. Science & Engineering
University of Florida
gmccune@ufl.edu

ABSTRACT

Internships provide opportunities for computing students to self-evaluate their interests and develop authentic technical and professional skills that are critical to a career in computing-related industries. However, it is a cause for concern that only 60% of computing students participate in an internship before graduation. Our work aims to identify the factors which are associated with the likelihood of a student's participation in an internship. To identify these factors, we designed a cross-sectional study at a large public university in the United States. 518 computing undergraduate students completed our survey, and we used a quantitative approach to model a student's ability to secure internships. Using a logistic regression model, we found that (1) year in school, (2) household income (a proxy for socioeconomic status), (3) involvement in activities outside the curriculum, and (4) lower identity diffusion scores (i.e., low exploration and low commitment) are significantly associated with a student's participation in an internship. Our findings confirm prior work which showed that factors outside the curriculum are at play for students' internship participation. Further, we add to the computing education research literature the unexplored relationship between computing students' identity formation and participation in internships.

CCS CONCEPTS

• **Social and professional topics**~Professional topics~Computing profession~Employment issues

KEYWORDS

internship, employment, identity, modeling

ACM Reference format:

Megan Wolf, Amanpreet Kapoor, Caroline Hobson, and Christina Gardner-McCune. 2023. Modeling Determinants of Undergraduate Computing Students' Participation in Internships. In Proceedings of 54th ACM Technical Symposium on Computer Science Education (SIGCSE '23), March 2023, Toronto, Canada. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3545945.3569754

SIGCSE 2023, March 15–18, 2023, Toronto, ON, Canada

ACM 978-1-4503-9431-4/23/03…$15.00
https://doi.org/10.1145/3545945.3569754

1 Introduction

Computer science (CS) employers have reported that computing graduates often lack technical and professional skills [11, 36]. Internships provide students with an opportunity to address this deficit and gain job-necessary skills before entering the workforce [14, 19, 24, 46, 48]. In addition, internships allow employers an opportunity to evaluate potential candidates, thus deeming internships crucial to the full-time recruitment process [33, 42, 45]. Therefore, encouraging students to participate in internships may be an effective strategy for preparing students for jobs in industry and reducing the skill-deficit. However, it is concerning that only 57.5% of the graduating computing students pursue an internship [23]. Our paper aims to elucidate the potential reasoning for the lack of internship participation by examining the factors that may influence computing students' ability to secure internships. We collected data from a single institution and used a binary logistic regression model to answer the following research question: *What are the factors that influence undergraduate computing students' participation in internships?* Our findings extend and confirm prior work and we contribute to computing education research (CER) literature the unexplored relationship between internship participation and identity formation. Our work can inform CS departments about student barriers to internship participation and aid in the development of support programs focused on improving students' employment outcomes.

2 Prior Work

Benefits of internships: why are they important?: Industry internships are a valuable method of gaining technical and professional skills beyond what is taught in formal academic computing classrooms. In addition, internships provide opportunities for self-evaluation, expansion of professional networks, and development of professional expectations [24, 31]. Thus, internships provide an authentic environment for addressing the industry-related skill deficit [36] among computing students. In addition to supporting students in skill building, internships have been shown to foster identity formation [8, 25], increase retention in computing programs [13], and improve capstone project quality [20, 34]. Lastly,

employers often utilize internships as evaluation tools during the recruitment process for hiring decisions, and students who previously have interned are more likely to get a full-time position and a higher starting salary [10, 33, 42, 45]. In summary, internships have been found to provide CS students with opportunities for skill building and participation in internships is beneficial for securing subsequent employment after graduation. Thus, it is highly recommended that students participate in internships before graduation.

Participation in internships in computing: Given the importance of internships, prior work has explored computing students' participation in these opportunities [23] and assessed if CS curricula prepares them for industry experiences [24, 47]. Our previous work explored the demographics of students who participate in internships and identified factors or barriers that support or prohibit students from securing internships using qualitative and bivariate analysis [22, 23]. Our work found that 57.5% of graduating seniors participate in an internship and students fail to participate in an internship because of low self-efficacy, alternate priorities such as job or family responsibilities, and application process challenges. In this work, we extend our prior work by using a multivariable regression model to determine the underlying factors that influence CS students' ability to secure internships.

Modeling and internships: Work that quantitatively modeled internship participation and factors includes a study by Hoekstra which modeled variables that predict internship participation across different undergraduate majors [18]. Her work found five significant pre-college predictors (race, gender, age, first-generation status, and future educational plans) and participation in high-impact practices such as research, learning communities, etc. influenced internship participation. According to her study, students who were Asian American, male, older, first-generation, or had lower participation in high-impact practices were less likely to have interned. Hoekstra's findings provide insight for building our model and we used several similar predictors. While Hoekstra's study generalizes across majors, we aim to extend this work by performing a similar analysis on the unexplored field of computing. Further, we include variables specific to CS curriculum, involvement, and identity status scales as predictors.

Internship participation has also been modeled in other fields including civil engineering [16] and medical sciences [15]. These studies determined potential predictors of students' satisfaction with internships [16] or modeled attributes that determine student success in internship performance [15]. Generically across majors, researchers have also predicted final grades and degree level classification from internship experience [3] or analyzed the impact of internships on university graduation rates [21]. In contrast, our study models the inverse of these relationships as we seek to understand if higher grades or other variables are predictors of securing internships.

Theoretical background on identity: James Marcia's theory of identity development [27, 29] operationalizes stages of identity development. Marcia's theory suggests that professional identity forms during ages 17-23 and identity changes over time-based on a person's active or passive exploration and commitment to their chosen profession or discipline. The theory identifies four statuses to characterize individuals' identity development: (1) *identity diffusion*, when an individual is neither exploring nor committed to a career choice; (2) *identity foreclosure*, when an individual has not explored career options but is committed to a career due to influence of an external agent; (3) *identity moratorium*, when an individual is exploring career options but is not committed to a career choice; and (4) *identity achievement*, when an individual has explored career options and is committed to an identity after the exploration process. The theory proposes that identity develops during active exploration highlighted by the moratorium and achievement statuses. Based on these statuses, we hypothesize that students in moratorium or achievement statuses are more likely to have interned than students in diffusion or foreclosure statuses. In an attempt to quantitatively represent identity statuses as predictor variables in our analysis, we used the validated Extended Objective Measure of Ego Identity Status (EOM-EIS) instrument [2] which is based on Marcia's theory.

Identity formation and internships: Prior work on identity formation and internships has found that internships support identity formation in engineering [8] and counseling psychology [9]. Our work tries to understand the inverse relationship between identity formation and participation in an internship. A study in Psychology by den Boer et al. [4] investigated the association between identity formation and internship participation and found that an internship in itself did not explain individual differences in identity processes, and enrollment in an internship was largely unrelated to identity processes, i.e. there is no relationship between internship participation and identity formation. den Boer et al.'s work is similar to our study, but the authors recruited graduate students who were interns and undergraduate students who did not intern and internship enrollment was an obligatory part of their curriculum. Our work pertains to the computing discipline, and we compare a more homogeneous population of undergraduate students who have or have not participated in internship(s). In addition, internship participation is optional in our program and hence our results might be incomparable with den Boer et al.

3 Methods

3.1 Study design

We designed a mixed methods study based on a Concurrent Triangulation Design, in which data is collected concurrently through multiple methods but is analyzed separately and later combined to triangulate overlapping patterns [6]. This design supports the corroboration of findings through multiple data sources and improves validity. We collected data through a cross-sectional survey and recruited students for interviews in 2019 after a pilot study in 2016 [24]. This study has a larger sample size (5x) compared to our pilot and findings from the pilot informed our study. For this paper, we focus on the survey

data and use a quantitative approach to model students' ability to secure internships. We aim to answer the following research question: *What are the factors that influence undergraduate computing students' participation in internships?*

3.2 Research Site

Our study population is traditional college students who are enrolled in an undergraduate CS-related major. Our sample is drawn from students enrolled in an undergraduate computing degree program at a large public university in the US. Admission to the site is selective and students can select a major when they start the program but have the flexibility to switch it at any time. Students in our sample were enrolled in CS, CE (Computer Engr.), and Digital Arts & Sciences (DAS) majors. Participation in an internship is not mandatory before graduation.

3.3 Participants and recruitment

Our study was approved by the Institutional Review Board at the research site. Participants were recruited from CS1, CS2, software engineering, HCI, and OS courses. The students were given 1% extra credit for participation. Alternatively, we offered gift cards to every 40th respondent if they chose to opt-out of extra credit. A substitute assignment requiring equal effort was also provided if a student did not wish to participate in our study. 698 students responded and consented to our survey after excluding 41 duplicates. The response rate was 43% (N=698, Total course enrollments=1620). From this dataset, the following were discarded: students who were not pursuing CS-related majors or were CS minors (n=78), students who completed less than 80% of the survey (n=20), students who were not in our undergraduate program (n=15), students without gender classification (n=2), non-traditional students over age 24 (n=45), and students with a high proportion of relevant missing data (n=20). Decisions to discard data were made for the following reasons: (1) we were trying to assess students' participation in internships who were enrolled in computing programs and represented traditional college students, (2) lack of data on a metric that was crucial for our analysis, and (3) inadequate representation of a certain population in our sample. Thus, our final corpus consists of 518 students who were enrolled in CS (66%), CE (26%), DAS (4%) or double (4%) majors. The average age of the respondents was 20.2 (Min=18, Max=24, SD=1.4). Other demographics are shown in Table 1 and are representative of the student population in CS program at our institution.

Table 1: Demographics of students in our corpus (N = 518)

Year					Gender		Race/Ethnicity				
1	2	3	4	5-6	M	F	White	Asian	Hispanic /Latinx	African American	Other
27%	18%	32%	17%	5%	73%	27%	47%	25%	20%	6%	3%

3.4 Data collection

Our study is a part of a larger project and our survey consisted of 11 sections (at most 74 questions, varied with display logic). Students spent 41.5 minutes on average completing the survey. The questions included 49 multiple-choice questions (MCQs), 10 short answers, and 15 open-ended responses. The sections spanned topics such as demographics, professional goals and identity, degree program experience, social support, and involvement in external activities. All questions were optional and were either developed from the findings of our pilot study [24, 26] or were taken from the following sources: NCWIT Student Experience of the Major Survey [43], CRA Data Buddies Survey [7], and the revised version of Bennion and Adams' Extended Objective Measure of Ego Identity Status (EOM-EIS) validated instrument [2] which consisted of measures to quantify Marcia's identity statuses [29]. In this paper, we use data from 16 questions which were selected for the following reasons: (1) our approach to analysis is quantitative and hence we discarded open-ended questions, (2) the question was irrelevant for answering our research question, and (3) the question provided background information on the sample or context for replication.

3.5 Response and explanatory variables

Our response (dependent) variable is a binary categorical variable representing participation in internship(s) or co-op(s) during a student's enrollment in a degree program (not counting internships during high school). We asked students if they had previously interned or were going to participate in an internship in the upcoming summer (they already received an offer). If the student answered yes to either of these choices, they were coded as "*yes*" as we are trying to understand students' ability to secure internships. We used 13 explanatory or independent variables (described in Table 2) in our model to identify their associations with a student's participation in an internship.

Eight of our explanatory variables were single-item measures such as Gender or GPA. The remaining five variables consisted of multiple-item measures. These multiple-item measures were aggregated to form scores representing four identity status variables and one variable called *External Involvement* which denotes a composite score for a student's involvement in activities outside the classroom such as hackathons, conferences, research, and student clubs, etc. For the *External Involvement* variable, we collected information on how frequently a student participated in activities outside the classroom using an ordinal scale for each activity ("Never {coded to 0}", "Once {1}", "2-3 times {2}" and "4 or more times {3}"). The composite score for each student was computed by aggregating the numerically coded responses of participation frequencies in all activities. For example, if a student stated that they participated in 2 of the 14 activities (e.g. personal projects and clubs), and they participated in each of them "Once" (coded as 1), their *External Involvement* score was 2 out of a maximum possible score of 42 (14 x 3).

Four MCQs in our survey that pertained to Marcia's identity statuses measured using the EOM-EIS instrument consisted of multiple items for an identity status (six 5-point Likert statements per status, 24 statements in total). For measuring each status, the scale included two statements that gauged identity status in relation to occupation, recreational activities, and lifestyle [2]. Thus, each status had a corresponding variable representing the aggregate of six ordinally coded 5-point Likert

Table 2: Explanatory (independent) variable descriptions

Variable Category	Independent Variable	Description (Coded value)
Demographic and Socioeconomic Factors	Household income ★	{"Less than $20,000" (1), "$20,000 to $34,999" (2), "$35,000 to $49,999" (3), "50,000 to $74,999" (4), "$75,000 to $99,999" (5), "$100,000 to $149,000" (6), "Over $150,000" (7)}
	Race/ethnicity ▲	{White/Asian (0), Underrepresented: all other ethnic and racial representations (1)}
	Gender ▲	{Male (0), Female (1)}
	Age ■	Numerical (Range: 18-24)
	Employment status ▲	{Unemployed (0), Employed - working along with the degree program (1)}
Academic Profile	GPA ■	University-level grade point average on a 4.0 scale
	High school courses in CS ▲	{No (0), Yes (1)}
	Year in school ★	{Freshman (1), Sophomore (2), Junior (3), Senior (4), Super Senior (5)}
Identity	Diffusion score ⊙	Marcia identity status composite score (scale: 6-30): Low exploration, low commitment
	Foreclosure score ⊙	Marcia identity status composite score (scale: 6-30): Low exploration, high commitment
	Moratorium score ⊙	Marcia identity status composite score (scale: 6-30): High exploration, low commitment
	Achievement score ⊙	Marcia identity status composite score (scale: 6-30): High exploration, high commitment
External Involvement	External involvement score ⊙	Composite score based on involvement in 14 activities, e.g. hackathons, clubs, etc.(scale: 0-42)

Key: Binary encoded categorical ▲ | Ordinal encoded categorical ★ | Quantitative ■ | Quantitative variable computed from ordinal scale questions ⊙

statements (Strongly disagree: 1 to Strongly agree: 5) and had a range between 6 - 30. A higher value in a status scale implies a higher likelihood for a student to be in that status.

We computed Cronbach's alpha, a measure of internal consistency of a scale that measures a latent variable (in our case each *identity status* was a latent variable). Cronbach's alpha measures how closely related a set of single-measure items are as a group. For our sample, Cronbach alpha coefficients were 0.64 for diffusion status, 0.83 for foreclosure, 0.61 for moratorium, and 0.62 for achieved status. While coefficient values of 0.70 or greater are an indicator of high reliability of an instrument in social sciences [44], Pallant argues that variables measured with less than 10 items generally have lower values of Cronbach's alpha [35]. For each status, we used six statements, and hence the lower values of the coefficient could be attributed to the lower number of items in our scale. Moreover, the range of values for our Cronbach's alpha coefficients was in line with the original EOM-EIS instrument [2] as well as subsequent studies that used this scale in other domains [5, 28, 39, 41]. Hence, there is a possibility that scales for measuring complex identity statuses have lower internal consistency than other constructs.

3.6 Data Analysis

We used a binary logistic regression model to identify consequential factors for securing internships. This model can be used to understand the relationship between categorical and continuous explanatory variable(s) and a dichotomous response variable [17]. The logit (i.e., the natural logarithm of an odds ratio, a measure that defines the ratio of successes to failures for an event) forms the basis of logistic regression. The odds ratio provides a measure that represents the odds that an outcome will occur (e.g., a student participates in an internship), given the presence of some other factor and controlling for other predictors. For example, we can obtain an odds ratio of a student's participation in an internship given they took high school courses in CS. This measure helps us quantify the strength of the correlation between demographic, academic, and identity factors and a student's participation in an internship(s).

The logistic regression equation takes the following form:

$$Z = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{12} X_{12}$$

where P_i is the probability of event i, β_0 is the constant coefficient, $X_1 \ldots X_{12}$ are the explanatory variables, and $\beta_1 \ldots \beta_{12}$ are coefficients of explanatory variables. A positive coefficient indicates a positive relationship between the response variable and the coefficient's respective explanatory variable, and a negative coefficient indicates a negative correlation. For a logistic regression model, the null and alternative hypotheses are as follows:

$$H_0 = \beta_1 = \beta_2 = \cdots = \beta_{12} = 0$$
$$H_A = \beta_1 = \beta_2 = \cdots = \beta_{12} \neq 0$$

In our model, we incorporate 12 explanatory variables after getting rid of one of the explanatory variables, *age*, due to multicollinearity (as described in Section 3.8). Therefore, our null hypothesis is that none of the predictor variables in our model have a statistically significant relationship with computing students' participation in internships. The alternative hypothesis is that at least one of the predictor variables in our model significantly contributes to CS students' probability of participating in internships. For our regression analysis, we treated ordinal explanatory variables (e.g., *year in school*) as continuous similar to other research in social sciences [37]. Our analysis including data cleaning and preprocessing was conducted in Microsoft Excel, IBM SPSS, and Python libraries such as pandas, matplotlib, seaborn, and researchpy [49]. Before feeding data into the model, we had to deal with missing data and multicollinearity which is discussed in the next subsection.

3.7 Data imputation and preprocessing

Our final data corpus consisted of missing data as all questions in our survey were optional. The total missing data for explanatory variables we imputed in this paper was 1.4% (n=326, N=23828 total data points). At a granular level, we imputed data (replaced missing values with substitute) for the following explanatory variables: GPA (n=19, N=518, 3.6%), age (n=43,

N=518, 8.3%), household income (n=47, N=518, 9.1%), identity achievement (n=5, N=3108 single-item measures, 0.2%), identity diffusion (n=9, N=3108, 0.3%), identity foreclosure (n=5, N=3108, 0.2%), identity moratorium (n=12, N=3108, 0.4%), and external involvement score (n=186, N=7252, 2.6%). We used imputation techniques depending on the type of the missing data (i.e., quantitative or categorical) and the skewness of a variable's distribution. It is recommended to not replace missing values with the mean for skewed data distributions because outliers are more likely to influence the mean, therefore we utilized either the median or mode [12]. We used the seaborn python library to plot a kernel density estimate and histogram with bins. We observed that the GPA distribution and household income were skewed towards the left, the external involvement score and age were skewed right, and identity scores were all slightly skewed (see Figure 1). As such we imputed missing values for numerical explanatory variables such as GPA, age, or scores with the median values and used the mode for household income which is a categorical explanatory variable.

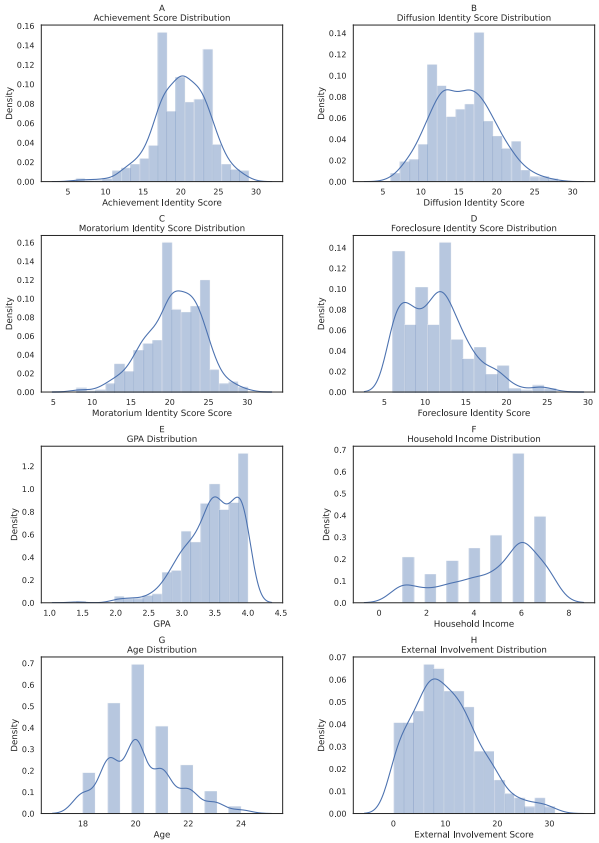

Figure 1: Distributions of imputed variables

Before conducting regression analysis, we also explored the possibility of correlations between our explanatory variables to limit the possibility of introducing multicollinearity in our model. Multicollinearity can lead to higher variance, overfitting, and difficulty in model interpretation due to instability in the magnitude of regression coefficients [17]. If a regression model is composed of two or more predictors that are moderately or highly correlated, multicollinearity exists. A common method for checking if multicollinearity exists in a model is checking for high correlations among pairs of predictor values. In our study, we check for correlations among the predictors using Pearson's R for continuous vs. continuous cases, Correlation Ratio for categorical vs. continuous cases, and Cramer's V or Theil's U for categorical vs. categorical cases using code modified from the dython library [40]. We consider correlation coefficients with a magnitude greater than ±0.7 to be highly correlated [32]. Also, we calculated the variance inflation factor (VIF) as an additional collinearity diagnostic metric [17]. A general rule is that if the VIF score for a predictor is greater than 5, it is recommended to remove one of the correlated explanatory variables to limit multicollinearity. The VIF values for predictors were below the threshold and ranged from 1.06 to 2.98. However, age was highly correlated with the year in school (Pearson's R = 0.80) and hence we excluded age as a predictor in our model.

4 Results

Our corpus consists of undergraduate students in computing majors enrolled at a single institution in the US (N=518). Trends in our explanatory variables and response variable are described in Tables 3, 4, and 5.

Table 3: Descriptive Statistics of Dependent/Response Variable

Variable (N=518)	Outcome	Count	%
Internship participation	No internship experience	301	58.1
	Participated in at least 1 internship	217	41.9

Table 4: Descriptive Statistics of Explanatory Ordinal/Numeric Variables

Variable (N=518)	Mean	SD	SE	95% Conf. Interval	
GPA	3.45	0.42	0.02	3.42	3.49
Achievement score	20.09	3.58	0.16	19.78	20.40
Diffusion score	15.54	4.02	0.18	15.20	15.89
Foreclosure score	11.44	4.04	0.18	11.09	11.79
Moratorium score	20.53	3.65	0.16	20.22	20.85
Involvement score	10.25	6.51	0.29	9.69	10.81
Household income	4.82	1.89	0.08	4.66	4.98
Year in school	2.54	1.20	0.05	2.44	2.65

Table 5: Descriptive Statistics of Explanatory Categorical Variables

Variable (N=518)	Outcome	Count	Percent
Secondary CS Education (High school CS courses)	No	261	50.39
	Yes	257	49.61
Employment status	Unemployed	360	69.50
	Employed	158	30.50
Gender	Male	379	73.17
	Female	139	26.83
Race	White or Asian	374	72.20
	Underrepresented	144	27.80

Our regression results can be found in Table 6 and our model is significant ($p < 0.001$). In this table, Coef. β represents the regression coefficient which estimates the relationship between the individual factor and whether students have received an internship. Std. Err represents the standard error, which measures the precision of the estimate of the coefficient. Z represents the Z-value, which is a test statistic that measures the ratio between a coefficient and the standard error. The Z-value is used to calculate the p-value of a factor, which is represented by $p > |z|$. A p-value is used to determine if a factor is statistically significance ($p < 0.05$). The odds ratio is a measure of practical significance and is represented by $\exp(\beta)$. If the odds ratio is greater than 1, the event that a student participates in an

internship is more likely to occur as the predictor value increases. The last two columns represent the confidence intervals, which show the range of values that the Odds Ratio could fall under with 95% confidence.

Table 6: Regression Results

	Coef. β	Std. Err	Z	p > |Z|	CI for Coef. β		Odds Ratio exp(β)	CI for Odds Ratio	
					[0.025	0.975]		5%	95%
Const	-5.53	1.60	-3.45	0.00	-8.67	-2.39	0.00	0.00	0.09
HS CS Edu.	0.29	0.21	1.36	0.18	-0.13	0.71	1.34	0.88	2.04
Employment	0.10	0.23	0.45	0.65	-0.35	0.56	1.11	0.70	1.75
Year in School	0.57	0.10	5.72	**0.00****	0.38	0.77	1.77	1.46	2.16
GPA	0.51	0.28	1.80	0.07	-0.05	1.06	1.66	0.96	2.89
Household Income	0.22	0.06	3.76	**0.00****	0.11	0.34	1.25	1.11	1.40
Gender	−0.28	0.24	-1.16	0.25	-0.76	0.19	0.75	0.47	1.22
Race	0.16	0.24	0.67	0.50	-0.31	0.63	1.18	0.73	1.88
Moratorium Score	0.01	0.03	0.48	0.64	-0.04	0.07	1.01	0.96	1.07
Diffusion Score	-0.06	0.03	-2.14	**0.03***	-0.12	-0.01	0.94	0.89	0.99
Achievement Score	0.01	0.03	0.29	0.77	-0.05	0.07	1.01	0.95	1.07
Foreclosure Score	-0.01	0.03	-0.34	0.74	-0.06	0.05	0.99	0.94	1.05
Involvement Score	0.12	0.02	6.79	**0.00****	0.09	0.16	1.13	1.09	1.17

No. of Observations: 518 | Pseudo R^2: 0.21 | LLR p-value: 4.7e-25
Df Residuals: 505 | Log-Likelihood: -279.5 | * p < 0.05
Df Model: 12 | LL-Null: -352.2 | ** p < 0.001

According to our regression results, *year in school*, *household income*, *identity diffusion score*, and *external involvement score* are significant. The odds ratio for the *year in school* indicates that for every one-year increase, a student is 1.77 times as likely to have participated in an internship, after controlling for other predictors. The odds ratio for *household income* indicates that for every one-unit increase (movement to the next socioeconomic status), a student is 1.25 times as likely to have participated in an internship (i.e., a one-unit increase in *household income* is associated with a 25% increase in the odds of a student participating in an internship). Similarly, the odds ratio for *external involvement* indicates that for every one unit increase in the external involvement score, a student is 1.13 times as likely to have participated in an internship. The odds ratio for *identity diffusion score* indicates that for every one unit increase in diffusion score, the odds of *not* securing an internship increases by a factor of 1.06 after controlling for other predictors [17].

Our model fit was evaluated using McFadden's pseudo-R^2 coefficient (ρ^2). The pseudo-R^2 of 0.21 indicates an excellent model fit. The values for pseudo-R^2 tend to be significantly lower than the standard R^2 and should not be interpreted by the same standards of fit as OLS regression. According to McFadden, "*values of .2 to .4 for ρ^2 represent an excellent fit*" [30].

5 Discussion and Conclusion

The regression results show that year in school, household income, external involvement score, and diffusion identity status score are significant predictors in our model. Therefore, we reject the null hypothesis that there are no factors in our model that have a significant relationship with computing students' participation in internships. These results corroborate the findings in our previous analysis [22, 23]. Our results also align with Hoekstra's study [18] which found age (correlated with the year in school) and participation in high-impact practices were significant predictors of securing internships in all majors. However, unlike Hoekstra's study, we did not find a relationship between race and gender and participation in internships in computing. Based on our findings, CS departments should provide resources for and underscore the importance of involvement in external activities. Moreover, because household income is a significant determinant of internship participation, universities should provide support to students coming from underprivileged backgrounds. Lastly, diffusion identity status seems to have a significant impact on internship participation. While higher exploration or higher commitment might not predict participation in an internship, a student in a lower commitment and lower exploration mode might face challenges in securing an internship. In the future, we would like to explore who are the students that are "stuck" in diffusion status. Are they freshmen or senior students? Finally, we would like to build on our model and explore creating predictive machine-learning models for predicting students' participation in internships in our future work.

6 Limitations and Threats to Validity

Our study has several limitations. First, data collected from surveys can induce response bias or interpretation of questions different from intended meaning of a prompt. Second, imputing missing data before running a model, has a chance to increase the underestimation of standard errors and overestimation of test statistics [1]. Given that the overall missing data was relatively low (1.3%) when compared with the number of responses that had missing data (40%), we decided to impute data rather than discard responses. We report our data imputation technique for better transparency. Third, our EOM-EIS identity scales had lower internal consistency due to the limited number of items used for each status. However, the Cronbach Alpha values were comparable to prior work. Fourth, we chose a logistic regression model for its simplicity, effectiveness, and lack of baselines. Prior work has observed that logistic regression produces somewhat comparable results as more advanced models in social sciences research [38]. However, our findings could be biased by the choice of our modeling technique. Fifth, our study is an observational study and results should not be interpreted as causal relationships. Finally, our data is from a modest sample of computing undergraduates enrolled at a single institution in the US where participation in internships was optional. The results may or may not generalize to other majors, institutions, or geographic areas, especially to programs where participation in internships is mandatory before graduation.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their valuable comments. This work is supported in part by the SIGCSE Special Project Grant. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the authors and do not represent the SIGCSE Board.

REFERENCES

[1] Allison, P. 2002. Missing Data. SAGE Publications, Inc.

[2] Bennion, L.D. and Adams, G.R. 1986. A Revision of the Extended Version of the Objective Measure of Ego Identity Status: An Identity Instrument for Use with Late Adolescents. Journal of Adolescent Research. 1, 2 (Apr. 1986), 183–197. DOI: https://doi.org/10.1177/074355488612005.

[3] Binder, J.F., Baguley, T., Crook, C. and Miller, F. 2015. The academic value of internships: Benefits across disciplines and student backgrounds. Contemporary Educational Psychology. 41, (2015), 73–82. DOI: https://doi.org/https://doi.org/10.1016/j.cedpsych.2014.12.001.

[4] den Boer, L., Klimstra, T.A. and Denissen, J.J.A. 2021. Associations between the identity domains of future plans and education, and the role of a major curricular internship on identity formation processes. Journal of Adolescence. 88, (2021), 107–119. DOI: https://doi.org/https://doi.org/10.1016/j.adolescence.2021.02.005.

[5] Cakir, S.G. 2014. Ego Identity Status and Psychological Well-Being Among Turkish Emerging Adults. Identity. 14, 3 (Jul. 2014), 230–239. DOI: https://doi.org/10.1080/15283488.2014.921169.

[6] Creswell, J.W. and Creswell, J.D. Research design: qualitative, quantitative, and mixed methods approaches.

[7] Data Buddies - Center for Evaluating the Research Pipeline: https://cra.org/cerp/data-buddies/. Accessed: 2019-08-31.

[8] Dehing, F., Jochems, W. and Baartman, L. 2013. Development of an engineering identity in the engineering curriculum in Dutch higher education: an exploratory study from the teaching staff perspective. European Journal of Engineering Education. 38, 1 (Mar. 2013), 1–10. DOI: https://doi.org/10.1080/03043797.2012.742866.

[9] Dunstan, L. 2002. Positioning and repositioning: Professional identity development during a counselling internship. Psychology in Society. 28, (2002), 40–47.

[10] Employers Prefer Candidates With Work Experience: http://www.naceweb.org/talent-acquisition/candidate-selection/employers-prefer-candidates-with-work-experience/. Accessed: 2018-08-31.

[11] Exter, M. 2014. Comparing educational experiences and on-the-job needs of educational software designers. Proceedings of the 45th ACM technical symposium on Computer science education - SIGCSE '14 (New York, New York, USA, 2014), 355–360.

[12] Feature Engineering Part-1 Mean/ Median Imputation. | by Arun Amballa | Analytics Vidhya | Medium: https://medium.com/analytics-vidhya/feature-engineering-part-1-mean-median-imputation-761043b95379. Accessed: 2022-08-13.

[13] Fryling, M., Egan, M., Flatland, R.Y., Vandenberg, S. and Small, S. 2018. Catch 'em Early: Internship and Assistantship CS Mentoring Programs for Underclassmen. Proceedings of the 49th ACM Technical Symposium on Computer Science Education - SIGCSE '18 (New York, New York, USA, 2018), 658–663.

[14] Gardner, J. and Veer, G. Van der 1998. The Senior Year Experience. Facilitating Integration, Reflection, Closure, and Transition. Jossey-Bass Inc., Publishers.

[15] Greenburg, D.L., Durning, S.J., Cohen, D.L., Cruess, D. and Jackson, J.L. 2007. Identifying Medical Students Likely to Exhibit Poor Professionalism and Knowledge During Internship. Journal of General Internal Medicine. 22, 12 (2007), 1711–1717. DOI: https://doi.org/10.1007/s11606-007-0405-z.

[16] Guler, H. and Mert, N. 2012. Evaluation of internship programs for educational improvements: a case study for civil engineering. The International journal of engineering education. 28, 3 (2012), 579–587.

[17] Hahs-Vaughn, D. and Lomax, R. 2013. An introduction to statistical concepts. Routledge.

[18] Hoekstra, C. 2021. Examining Demographic and Environmental Factors that Predict Undergraduate Student Participation in Internships. (2021).

[19] Internships Enhance Student Research and Educational Experiences: https://cra.org/crn/2008/11/internships_enhance_student_research_and_educational_experiences. Accessed: 2019-08-31.

[20] Jaime, A., Olarte, J.J., García-Izquierdo, F.J. and Domínguez, C. 2020. The Effect of Internships on Computer Science Engineering Capstone Projects. IEEE Transactions on Education. 63, 1 (2020), 24–31.

[21] Johnson, S.R. and Stage, F.K. 2018. Academic Engagement and Student Success: Do High-Impact Practices Mean Higher Graduation Rates? The Journal of Higher Education. 89, 5 (Sep. 2018), 753–781. DOI: https://doi.org/10.1080/00221546.2018.1441107.

[22] Kapoor, A. and Gardner-McCune, C. 2020. Barriers to securing industry internships in computing. ACE 2020 - Proceedings of the 22nd Australasian Computing Education Conference (2020).

[23] Kapoor, A. and Gardner-McCune, C. 2020. Exploring the Participation of CS Undergraduate Students in Industry Internships. SIGCSE 2020 - Proceedings of the 51st ACM Technical Symposium on Computer Science Education (2020).

[24] Kapoor, A. and Gardner-McCune, C. 2019. Understanding CS undergraduate students' professional development through the lens of internship experiences. SIGCSE 2019 - Proceedings of the 50th ACM Technical Symposium on Computer Science Education (Feb. 2019), 852–858.

[25] Kapoor, A. and Gardner-McCune, C. 2019. Understanding CS undergraduate students' professional identity through the lens of their professional development. Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE '19 (Jul. 2019), 9–15.

[26] Kapoor, A. and Gardner-McCune, C. 2018. Understanding Professional Identities and Goals of Computer Science Undergraduate Students. Proceedings of the 49th ACM Technical Symposium on Computer Science Education (New York, NY, USA, 2018), 191–196.

[27] Kroger, J. and Green, K.E. 1996. Events associated with identity status change. Journal of Adolescence. 19, (1996), 359–380.

[28] Lewis, T.F. 2006. Discriminating Among Levels of College Student Drinking Through an Eriksonian Theoretical Framework. Journal of Addictions & Offender Counseling. 27, 1 (2006), 28–45. DOI: https://doi.org/10.1002/j.2161-1874.2006.tb00016.x.

[29] Marcia, J.E. 1966. Development and validation of ego-identity status. Journal of Personality and Social Psychology. 3, 5 (1966), 551–558. DOI: https://doi.org/10.1037/h0023281.

[30] McFadden, D. 1977. Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments.

[31] Minnes, M., Serslev, S.G. and Padilla, O. 2021. What Do CS Students Value in Industry Internships? ACM Trans. Comput. Educ. 21, 1 (Mar. 2021). DOI: https://doi.org/10.1145/3427595.

[32] Moore Notz, William., Fligner, Michael A., D.S. 2013. The basic practice of statistics. W.H. Freeman and Co.

[33] National Association of Colleges and Employers 2014. The Class of 2014 Student Survey Report.

[34] Olarte, J.J., Dominguez, C., Jaime, A. and Garcia-Izquierdo, F.J. 2019. Impact of Part-Time CS Engineering Internships on Workload. Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education (New York, NY, USA, 2019), 318.

[35] Pallant, J. 2002. SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS. (Jul. 2002), 378. DOI: https://doi.org/10.4324/9781003117452.

[36] Radermacher, A. and Walia, G. 2013. Gaps between industry expectations and the abilities of graduates. Proceeding of the 44th ACM technical symposium on Computer science education - SIGCSE '13 (New York, New York, USA, 2013), 525.

[37] Robitzsch, A. 2020. Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods. Frontiers in Education. 5, (2020). DOI: https://doi.org/10.3389/feduc.2020.589965.

[38] Salganik, M.J. et al. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. Proceedings of the National Academy of Sciences of the United States of America. 117, 15 (Apr. 2020), 8398–8403.

[39] Schwartz, S.J. 2006. Predicting identity consolidation from self-construction, eudaimonistic self-discovery, and agentic personality. Journal of Adolescence. 29, 5 (2006), 777–793.

[40] shakedzy/dython: A set of data tools in Python: https://github.com/shakedzy/dython. Accessed: 2022-08-13.

[41] Sharma, T. and Mittal, U. Identity Diffusion:Role of Parenting Style and Decision Making Style among Adolescents. Indian Journal of Health and Wellbeing; Vol 8, No 7 (2017).

[42] Stepanova, A., Weaver, A., Lahey, J., Alexander, G. and Hammond, T. 2021. Hiring CS Graduates: What We Learned from Employers. ACM Trans. Comput. Educ. 22, 1 (Oct. 2021). DOI: https://doi.org/10.1145/3474623 .

[43] Student Experience of the Major (SEM): https://www.ncwit.org/sites/default/files/resources/sem_survey_in_a_box_0.pdf. Accessed: 2018-01-21.

[44] Taber, K.S. 2018. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Research in Science Education. 48, 6 (2018), 1273–1296. DOI:https://doi.org/10.1007/s11165-016-9602-2.

[45] The Positive Implications of Internships on Early Career Outcomes: http://www.naceweb.org/job-market/internships/the-positive-implications-of-internships-on-early-career-outcomes/. Accessed: 2018-08-31.

[46] Thiry, H., Laursen, S.L. and Hunter, A.-B. 2011. What Experiences Help Students Become Scientists? A Comparative Study of Research and other Sources of Personal and Professional Gains for STEM Undergraduates. The Journal of Higher Education. 82, 4 (Jul. 2011), 357–388.

[47] Tomer, G. and Mishra, S.K. 2016. Professional identity construction among software engineering students: A study in India. Information Technology and People. 29, 1 (2016), 146–172. DOI: https://doi.org/10.1108/ITP-10-2013-0181.

[48] Zehr, S.M. 2016. Student Internship Experiences And Learning Opportunities: A Mixed Methods Study.

[49] Wolf, M., Kapoor, A., Hobson, C., & Gardner-McCune, C. 2022. Modeling Determinants of Undergraduate Computing Students' Participation in Internships (Version 1.0.0). https://github.com/kapooramanpreet/Modeling-Internships