Programming, Research and... Coffee? An Analysis of Workplace Activities by Computing Interns

Huib Aldewereld

Institute for ICT

HU University of Applied Sciences Utrecht

Utrecht, the Netherlands

huib.aldewereld@hu.nl

Esther van der Stappen

Research Centre for Learning and Innovation

HU University of Applied Sciences Utrecht

Utrecht, the Netherlands

esther.vanderstappen@hu.nl

ABSTRACT

To overcome the skills gap between industry demands and learning outcomes achieved by graduates in higher computing education, many Bachelor programs integrate some form of internship in their curriculum; students are assumed to encounter authentic tasks and recent technologies in the workplace. In practice, however, educators often do not know specifically which tasks their students perform and which technologies they use, mainly due to the distance between coach and student during the work placement and the lack of cohort-based overviews of activities performed in internships. In this study, we gathered and analyzed workplace activity data of 54 students over the course of their third-year semester-long internships in the computing industry. We performed descriptive analyses to gain insight into i) which categories of activities students performed most (programming, research and documentation) and ii) which of the activity categories they find most difficult (research, documentation (both academic and IT) and implementation/configuration). Subsequent text analysis gives us insight into students' perceptions of the categories used to label activities (testing, research, meetings, and academic documentation are congruous) and which technologies were used most by these students. Based on the results, we conclude it is feasible to use user-generated data to get insights into workplace activities of computing interns. The quality of this user-generated data does hamper us in drawing certain conclusions. Further research is needed with improved data quality and volume in order to obtain more generalizable results.

CCS Concepts

• **Social and professional topics** → **Professional topics** → **Computing education** → Computing education programs • **Social and professional topics** → **Professional topics** → **Computing profession** → Employment issues • **Applied computing** → Education

Keywords

Computing Internships; Workplace Learning Analytics; Data-Driven Curriculum Development

ACM Reference format:

Huib Aldewereld and Esther van der Stappen. 2019. Programming, Research and... Coffee? An Analysis of Workplace Activities by Computing Interns. In *Proceedings of 8th Computer Science*

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

CSERC '19, November 18–20, 2019, Larnaca, Cyprus

© 2019 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-7717-1/19/11.

<https://doi.org/10.1145/3375258.3375264>

Education Research Conference (CSERC'19). ACM, Larnaca, Cyprus. <https://doi.org/10.1145/3375258.3375264>

1. INTRODUCTION

To overcome the skills gap between industry demands and learning outcomes achieved by graduates in higher computing education, many higher education computing programs integrate some form of workplace learning in their curriculum. The rationale is that internships better prepare students for their entry into the workplace after graduation, due to the authentic tasks and recent technologies that students encounter while working and learning in industry. Furthermore, learning in the workplace allows students to transfer computing knowledge and skills from university to professional contexts more easily.

In practice however, educators often do not know specifically which tasks their students perform in the workplace and which technologies they use. This lack of insight is mainly due to the distance between coach and student during the work placement. Generally, only a few visits take place where the coaching teacher visits the workplace of the student. Higher educational institutes also experience a lack of cohort-based overviews of activities performed in internships, since every single teacher only coaches a small group of interns and a broader collection of workplace learning data is rare.

Learning analytics endeavors in higher education have primarily focused on classroom-based learning. Recently, workplace learning analytics (WPLA) has become an emergent research area. WPLA opens up opportunities to analyze workplace learning data of students in computing programs in higher education and give us more insight into which knowledge and skills are necessary for computing students in their internships and hence should be covered in the curriculum beforehand. Especially for higher professional education – as is offered mainly at universities of applied sciences – workplace learning analytics seems to be crucial in data-driven or data-informed curriculum development.

2. THEORETICAL BACKGROUND

In this section, we discuss the theoretical background of this study focusing on the skills gap, workplace learning and technology-enhanced learning. Finally, based on this overview, we identify the goal of this study.

2.1 Skills Gap and Workplace Learning

Due to the ever-faster digitization of our society, our workplaces and our everyday life, the need for well-equipped graduates in the IT and computing domain has risen sharply [10]. According to the EU, there is a significant 'Digital Skills Gap' in Europe: business find it hard to employ both IT specialists as well as non-IT personnel with sufficient digital literacy [6, 7]. To fulfill the needs imposed by the digital transformation in Europe, two problems need to be tackled: 1) a shortage of computing graduates and 2) a

mismatch between graduates' qualifications and industry requirements. In this study, we focus on the latter.

In comparison to other domains, computing graduates need a very diverse skills set, ranging from hard skills, to innovative skills to social skills [25]. In a study among recent computer science graduates and employers of these graduates, a skills gap was found to exist for many of the skills that employers require of new employees [27]. In a large study among Spanish employers, it was found that in general, employers feel that graduates lack field-specific practical knowledge, whereas the gap for field-specific theoretical knowledge is much smaller [9].

Workplace learning within formal higher education programs is believed to help overcome the skills gap between graduates' qualifications and industry requirements. Gaining work experience during a Bachelor study offers students an opportunity to develop both 'hard skills' (subject-matter) and 'soft skills' such as teamwork, planning and communication [11]. Graduates with prior work experience are generally considered to have a higher 'employability' [2].

Workplace learning offers students opportunities to practice job-specific functions, such as socialization, innovation and job performance [16]. An increasing number of universities is integrating some form of workplace learning into their curricula. The pedagogy of such integrations [26] and the design of such hybrid learning environments [29] have been the subject of recent studies.

2.2 Technology in Education

Technology-Enhanced Learning (TEL) refers to the use of technology to support learning and teaching. TEL is often used as a synonym for e-learning but can also be used to refer to technology enhanced classrooms and learning with technology, rather than just through technology [23]. Besides e-learning, other examples of enhancing learning through the use of technology are Massive Open Online Courses (MOOCs) and Computer-Supported Collaborative Learning (CSCL). TEL provides the advantage of easier access to more information and creates flexibility in time and location of learning. Generally speaking, TEL can be used to (i) replicate existing teaching practices, (ii) supplement existing teaching or (iii) transform teaching and/or learning processes and outcomes [12].

Most studies that consider technology to support learning in the workplace, focus on the (formal or informal) learning of professionals working in industry and governmental organizations, often from a human resources point of view [19]. Examples of such technologies are (personal) knowledge management tools, e-learning platforms, mobile learning applications and social collaboration systems. Only few studies aim to design, develop and evaluate technologies that specifically support workplace learning, which is informal, contextual and social in nature [20, 22].

Recently, design propositions for Technology-Enhanced Workplace Learning (TEWL) have been developed and evaluated, and a web application was developed that implemented most of these propositions [22]. This open-source web application provides students with an interface to register their working and learning activities in the workplace in an easy-to-use way, which in turn allows for analytics (a dashboard with charts) giving them insight into their learning process.

A specific research area within TEL is Learning Analytics (LA), which provides data-driven interventions in the learning process, such as LA dashboards and predictive or recommendation

functionalities within a Learning Management System (LMS). The most commonly used definition of learning analytics is the one presented at the first Learning Analytics & Knowledge (LAK) conference: 'Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs' [1].

Learning analytics for workplace and professional learning is a recent area of interest. In 2016, a workshop on this topic was organized at the LAK conference [14]. One year later, a first systematic literature review on workplace learning analytics was published [17]. Only a few of the studies included in this review considered workplace learning as part of formal (higher) education. In a recent, extensive literature review of LA literature, the number of studies gathering data from informal settings (e.g. the workplace) and having a program-wide scope turned out to be very small, and even decreasing over the last few years [4].

Over the last few years, several approaches have been developed to improve education programs and curricula in a data-driven way. Data-Based Decision Making (DBDM) mainly uses hard data such as test scores to guide instructional decisions such as adaptation of teaching strategies [13, 28]. Schildkamp and Kuiper published a comprehensive overview of the use of data to inform curriculum development in secondary education [18]. For higher education, in 2014, an integral approach (including a process and the design of a tool) was proposed to enable data-driven course design [5]. Specifically for higher computing education, Méndez et al. presented an overview of techniques to analyze school-based data to obtain recommendations for curriculum re-design [15] and Toetenel and Rienties analyzed learning designs of university courses in order to evaluate the impact of pedagogical decision making [24].

Research on analyzing data collected in the workplaces of student internships in the computing domain is rare. Recently, possibilities were explored to use the data generated by the users of a TEWL application for data-driven curriculum development and requirements were elicited by evaluating the proposed solution with program managers [21].

2.3 Research Goal

For (computing) programs in higher professional education wishing to improve their curriculum in a data-driven way, it is crucial to analyze workplace data from student internships. In this study, we aim to answer the following three research questions.

When students in higher professional computing education perform an internship in industry:

- *Which activities do they perform most?*
- *Which activities do they find most difficult?*
- *Which technologies do they use most?*

The answers to these questions help to open the 'black box' of internships in higher computing education, which in turn can help universities to (re-)design their curriculum to better prepare students for professional practice.

3. METHODOLOGY

3.1 Participants

In the Fall Semester of 2018, we deployed and configured the aforementioned open-source TEWL application [22] and provided it as an opt-in instrument for a group of 183 third-year students in three Bachelor programs in the IT and Computing domain in a European university of applied sciences. It was mandatory for all students to record their worked hours; they were offered a choice

Figure 1: Screenshot of input screen of the application (with categories customized for the SE cohort).

between their own logbook solution (e.g. in Word or Excel) or to use our application. Out of the 183 students, 81 registered an account in our application, of which 68 entered at least one activity. We note that the participating students were self-selected, which might give rise to biased data.

Students performing an internship in these Bachelor programs have to spend at least 100 working days of at least 7.5 hours within an organization in the IT industry. Participants that registered less than four weeks of activities (equal to 20 working days, ca. 160 working hours) were removed from the data set, in order to have data representative of working activities for a significant amount of time with respect to the default internship of 100 days. This yielded a final set of 54 students (49 male, 5 female) that were included in this study. The average number of days per student on which activities were registered is 90 ($\sigma=25.2$ days).

All participants gave informed consent to the use of the registered data for research purposes. All data was anonymized before analysis.

3.2 Data Collection

In the application we provided to the participants, students register their working activities in an input screen as shown in Figure 1. The relevant part of the underlying database scheme is shown in Figure 2.

Central to the database model of the application is the *LearningActivity* table of which the attributes are recognizable in the input screen of the application. A student enters a description of an activity performed during their internship and labels this with (meta)data such as a date, a category, resources used, a status and a difficulty.

For authentication, administration and customization purposes, the tables *Internship*, *Student*, *EducationProgram* and *Cohort* are included. A *Cohort* is a group of students enrolled in the same program at the same time (e.g. all Software Engineering students starting their internship in September 2018). For our study, the table *Category* is interesting as this is used to label activities into specific classes that should give insight into the nature of these activities.

Categories can be defined i) at the education program level, ii) at the cohort level or iii) user-generated by a student. For this study, all categories predefined by us were defined at the cohort-level and

Figure 2: Core of the Entity-Relationship Diagram (ERD) of the application.

based on a national framework aimed at defining learning outcomes for Bachelor programs in the IT/Computing domain.

3.3 Analysis

We started our analysis with the necessary data extraction and data cleaning.

3.3.1 Descriptive analytics

As can be seen in Figure 1 and 2, students input data into a single text field (*Description*), after which they label the activity with meta data: *Duration*, *Category*, *Status*, *Difficulty* and optionally a previous activity. In our analysis, we do not use this chaining functionality and we also neglect the associated *Status* field (activities labelled as ‘Busy’ become available for chaining later activities).

We filtered all activities of the 54 participants from the MySQL database and exported them as comma-separated files for further analysis. We first explored the workplace data by computing several statistics on the number of activities, the number of days and the total duration of the registered activities per student; see Section 4.1. After this, the main focus of the descriptive analysis is on the *Category* field. We analyzed which categories were i) user-

Figure 3: Text analysis pipeline.

generated (Section 4.2), ii) most occurring (Section 4.3) and iii) most difficult (Section 4.4).

3.3.2 Text analysis

After the descriptive analyses of the meta-data of the entered activities, we performed a content analysis of the *Description* field. The analysis was focused on trying to determine the intent of the description as written by the student, by looking for commonly occurring keywords. This was done for each of the pre-defined categories of activities (*Research*, *Programming*, *Implementation/configuration*, *Testing*, *Academic Documentation*, *IT Documentation*, and *Meeting*). The descriptions were analyzed for each separate category, for all cohorts combined. The hypothesis was that the descriptions of the activity would largely match the category of the description (e.g., most students would mention something related to *Programming* in the description of a *Programming* activity), and that this would be no different between the cohorts.

To be able to check the descriptions for frequently occurring keywords, a Dutch Natural Language Processing toolkit called LaMachine¹ was used. In particular, a morphological tagger and parser named FROG [3] was used to tokenize and lemmatize the words (i.e., to make sure that “program”, “programming” and “programmed” are counted as a single keyword). Furthermore, we analyzed the descriptions on frequently occurring patterns of words (called n-grams) by means of the COLIBRICORE tool [8] also part of LaMachine.

The text analysis pipeline applied is shown in Figure 3. Results of this analysis are presented in Section 4.5.

A second text analysis was performed by combining all descriptions of the *Programming*, *Implementation/Configuration*, *Testing*, and *IT Documentation* activities, but separated by Cohort. These sets of descriptions were analyzed again using the same text analysis pipeline, but subsequently filtered for nouns, verbs and special named words. The resulting set of keywords was manually filtered on mentioned technologies. Each of these mentioned technologies was then used to determine the number of students mentioning using that technology (that is to say, instead of counting the frequency of that technology occurring in the descriptions, we counted the number of distinct students mentioning that technology instead). This gives us an indication of the most frequently used technologies. The results of this analysis are shown in Section 4.6.

4. RESULTS

4.1 Descriptive Statistics

The participants were students of three different Bachelor’s programs (cohorts): Business Informatics (BI), Software Engineering (SE) and IT Systems & Networks (SN). In Table 1, descriptive statistics (mean and standard deviation) of the data on the workplace activities entered by the participants is shown: the number of activities registered, the total number of hours spent on all activities in an internship, the number of days with registered activities, the number of activities registered per working day and the duration of a single activity (in hours).

Out of the 54 students included in this study, six students entered activities with an average duration of 7 hours or more. Of these six, four students consistently entered a single activity per working day (average duration of a single activity $\approx$ 8h). We observe that even though some students might register merely a single or two activities per day, in professional practice it is seldom the case that one performs a single same activity for a longer period of time (and that for many days in a row). We presume these students did not want to put in the (extra) effort of entering several activities per day, even if this means they will not get usable insights from the analytics of these data.

As a measure of variety in types of activities performed in the workplace, we analyzed how many different categories the participants used to label their registered activities; see Figure 4.

Figure 4: Number of different categories used by students to label their activities in the application.

¹ <https://proycon.github.io/LaMachine/>

The median is eight different categories. Notably, one student only used a single category to label all activities performed during their internship, whereas another student used as much as thirteen different categories. We notice the BI students on average used the least number of categories (6.2), whereas SE students used most (8.2).

4.2 User-generated Categories

As mentioned earlier (Section 3.2), a set of categories was pre-defined for each cohort, based on the learning outcomes of the Bachelor programs:

Table 1. Descriptive statistics of the workplace learning activities entered by the participants (per student).

Cohort	#students	#activities		total duration in hours		#days		#activities per day		duration of a single activity in hours	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
BI	12	195	118.0	752.8	156.00	94.6	19.53	2.1	1.09	4.9	2.18
SE	29	204	93.5	735.6	161.08	92.5	19.40	2.2	0.90	4.0	1.23
SN	13	150	88.8	635.1	278.51	79.8	36.10	1.9	0.64	4.8	1.67
Total	54	189	100.9	715.2	200.24	89.9	25.18	2.1	0.90	4.4	1.64

- All cohorts:
 - o Academic Documentation
 - o IT Documentation
 - o Meeting
 - o Research
 - o Testing
- BI & SN:
 - o Implementation / Configuration
- SE:
 - o Programming

These pre-defined categories were the categories most-used by students to classify their activities. However, the application also allows users to create their own categories. We presume that students that add customized categories feel that the pre-defined categories do not cover some of the activities they performed during their internship. We analyzed which categories were added by the computing students by extracting those from the database and classifying them into classes of activity categories. The results are presented in Table 2.

Notably, sixteen students created an ‘Other’ category for activities that did not fit the pre-defined categories. Moreover, five students created a ‘Container category’ with customized labels such as ‘Internship’, ‘Internship activities’, or ‘Work’. Of course, these categories do not give much insight into which specific activities are performed by our students in the workplace. Other user-generated categories that were added while they do not provide additional insight into the work activities of computing students are categories related to the category classes ‘Illness’, ‘Break / Free time / Drinks’, ‘Absent’ and ‘Outside of work’ (or when the category label was left empty). It is interesting that students did actually register activities in such categories, since absent days and activities outside of work do not count as working hours according to academic guidelines at the university of the participants. In Table 2, all user-created categories that do not give any new insights into

what students do are shaded dark grey.

The category classes that are shaded light grey actually overlap with pre-defined categories: Research, Meeting and (IT) Documentation. The students that created these categories may have overlooked the corresponding categories that were already available.

All categories that are not shaded grey in Table 2, do give us more insight into students’ working activities. Most notable are the category classes ‘Introduction’, ‘Training / course / personal development’, ‘Demo / presentation’ and ‘Setting up (technical)

Table 2. Student-generated activity categories.

Category class	BI	SE	SN	Total #students
Other	4	8	4	16
Introduction	1	5	7	13
Illness	1	5	5	11
Training / course / personal development	4	5	1	10
Demo / presentation	2	6	1	9
Preparation	1	5	2	8
Setting up (technical) working environment		4	2	6
Container category	1	1	3	5
Planning		5		5
Break / Free time / Drinks	1	2	1	4
Scrum activities		3	1	4
Teambuilding		4		4
Research	3	1		4
University work		4		4
Meetings	1	2	1	4
Assisting others			3	3
Absent	1	2		3
(IT) Documentation	1	1		2
-left empty-	1	1		2
Proof of concept	2			2
External visit / meeting		1	1	2
Problem solving		2		2
Organizing		1		1
Administration	1			1
Deployment		1		1
Outside of work		1		1

working environment’. Often, the internship is the first industry experience for a student, so it makes sense that many students register activities dealing with getting to know the company and the assignment, getting familiar with and setting up their working environment (such as a DTAP-pipeline) and taking in-house training or courses.

Only seven user-generated categories were used to label more than 100 working hours of a student, of which three were a ‘container category’. The other four were: ‘Documentation’, ‘Research’, ‘Gathering information’ (classified as Introduction) and ‘Proof of Concept’.

4.3 Most Occurring Activities

To find the most occurring activity categories, we summed up the durations of all activities and grouped them by category; we merged the user-generated categories Research, (IT) Documentation and Meeting with their corresponding pre-defined categories. We disregarded those categories that had a total duration of less than 100 hours over all 54 students. In Table 3, the relative occurrence of the remaining activity categories can be found. Since we did not include all categories, the percentages do not necessarily add up to 100%. In the right-most column, the relative occurrence of the category over students of all three cohorts is shown.

Table 3. Relative occurrence of activity categories over the three different cohorts.

Category class	BI	SE	SN	Overall
Programming		44.7%		25.7%
Academic documentation	29.1%	16.4%	32.3%	22.4%
Research	27.5%	13.6%	21.1%	18.2%
IT Documentation	13.2%	13.6%	12.6%	13.3%
Implementation/ configuration	8.1%		19.1%	5.7%
Meeting	5.7%	4.2%	4.4%	4.6%
Container category	8.4%	0.2%	4.3%	2.8%
Testing	1.1%	2.7%	2.3%	2.2%
Other	1.1%	0.5%	0.8%	0.7%
Introduction	1.7%	0.3%	0.6%	0.7%
Illness	0.3%	0.5%	0.9%	0.6%
Planning		0.8%		0.5%
Preparation	0.2%	0.5%	0.3%	0.4%
Proof of concept	1.6%			0.4%
Training / course / personal development	0.3%	0.4%	0.0%	0.3%
Demo / presentation	0.8%	0.1%	0.1%	0.3%

From Table 3, we can conclude that the pre-defined categories as listed in the previous section, were the categories that were used most by students (the sum of their percentages is 92.1%). The ‘Container category’ class is the most-used user-generated category, which is not surprising since students presumably use such categories to label all of their activities.

We also notice that SE students spend almost half of their time on programming, while BI and SN students clearly spend less time on the corresponding category Implementation / configuration. In turn, BI and SN students spend over half of their time on the two categories Academic Documentation and Research, whereas for SE students, this combination of categories accounts for less than a third of the total time spent. Regarding IT Documentation, Meeting and Testing, all cohorts relatively spend the same amount of time: approximately 13%, 5% and 2% respectively.

For the user-generated categories, we believe the most notable insight is that BI students spend more time on both Introduction and Demo / presentations, as compared to SE and SN students.

Figure 5: Ten most occurring categories for students from the program Business Informatics (BI) in total number of hours.

Figure 6: Ten most occurring categories for students from the program Software Engineering (SE) in total number of hours.

Figure 7: Most occurring categories for students from the program IT Systems & Network (SN) in total number of hours.

In Figures 5, 6 and 7, we picture the ten most occurring categories for each of the three cohorts separately for a more visual indication of the distribution of hours spent on the various activity categories.

Table 4. Distribution of difficulty levels over all registered activities.

Difficulty	Activities	Percentage	Duration
Easy	6,101	59.8%	20,419h
Average	3,394	33.2%	14,906h
Difficult	714	7.0%	3298h
Total	10,209	100%	38,623h

4.4 Difficulty of Activities

As shown in Figure 1, students can label the activities they register with the perceived difficulty. We allocated weights to these difficulty levels as follows: Easy=1, Average=2 and Difficult=3.

First, we studied some descriptive statistics of the difficulty labels entered by students; see Table 4. We note that a mere 7% of the activities is labelled Difficult by the participants, whereas almost 60% of activities is labeled Easy.

The original goal of the TEWL application we used in this study, is to create more awareness of the learning process and to realize more of the learning potential of students working in internships [20], e.g. by showing a reflection prompt (with feedback and feedforward questions) whenever students enter an activity with the difficulty level Average or Difficult and with the Status Busy. We hypothesize that students are generally not motivated to fill out a reflection prompt and as a reaction will label their activities as Easy, leading to only a small percentage of activities labelled Difficult. Further in-depth and qualitative analysis can shed light on whether students indeed find most of their work during internships as easy as they have inputted in the application.

Next, we analyzed the filled-out reflection prompts to gain insight as to why students found specific activities difficult. A vast majority of students (83%) indicated that the reason they perceived that activity difficult was a lack of experience. A mere 3% of the filled-out prompts recorded that a lack of resources was the culprit. Finally, we analyzed which activity categories were perceived to be most difficult. Since the default categories were the most occurring, we performed our analysis on these categories. We grouped all activities into the three difficulty levels and computed relative frequencies based on three aspects:

- the total number of activities (Figure 8);
- the total duration of the activities (Figure 9);
- the number of different students (Figure 10).

The reason we chose to perform three different analyses is that every single one says something about the difficulty of a category, but none of them says it all. The simplest measure is the first one (the total number of activities) since it gives insight into the dispersion over the difficulty classes within a given category. However, it does not take into account the duration. Is a difficult activity of 15 minutes more difficult than a difficult activity of 5 hours? Maybe the activity took so long just because it was so difficult? There is no straightforward solution, therefore we include both analyses. Moreover, the number of different students that labelled at least one activity in a specific category as difficult also gives insight, because a single student can increase the difficulty of a specific category significantly if she finds activities in the same category difficult often and these activities have high duration. Note that for this last analysis, students can register multiple activities within a category with potentially different difficulty labels. In Figure 10, percentages are computed relative to the number of students that registered activities within a category. Thus, the percentages per category add up to more than 100%.

Figure 8: Relative frequency of difficulty labels based on number of activities. **Figure 9: Relative frequency of difficulty labels based on duration.** **Figure 10: Relative frequencies of difficulty labels based on number of students.**

In Figure 8, we observe that Academic Documentation and Research have the highest relative frequencies for the label Difficult, whereas Meeting and Programming have the lowest. In

Figure 9, where we have taken duration into account, we see that Implementation/ configuration has the highest relative frequency for the label Difficult. For the other categories, the relative frequency for the label Difficult stays very similar to Figure 8. Notably, Academic documentation, in both Figure 8 and 9, is the only category for which the percentage for Easy is less than the percentage for Average. Finally, in Figure 10, we see that Academic and IT documentation have the highest difficulty. The other categories do not give surprising results.

From these three analyses, we can conclude that Academic documentation, Research, Implementation/ configuration and IT Documentation are perceived as most difficult. Meeting clearly emerges as the easiest category.

4.5 Students' Perceptions of Activity Categories

By means of the text analysis pipeline shown in Figure 3, we analyzed the descriptions entered by the students for each of their activities. We analyzed whether the descriptions were congruous to the activity's category. The analysis described in Section 3.3.2 was performed on the original Dutch texts, the keywords were translated to English only for presentation in this paper.

The descriptions of the activities in the following categories were inconclusive:

- *Implementation/Configuration*: 'implement' was mentioned in only 9.8%, and 'configure' was mentioned in only 8.8% of all 510 descriptions;
- *Programming*: 'program' was mentioned in only 5.1% of 2140 descriptions²; and
- *IT Documentation*: 'documentation' was mentioned in 8.1% of 800 descriptions).

Each of these keywords mentioned was also the highest ranking keyword for that category.

The results from the analysis of the categories *Meeting*, *Testing*, *Research*, and *Academic Documentation* are shown in Figure 11 to 14. Each of these categories has a keyword used in the descriptions of the activity congruous to the activity to which it is related. In the *Meeting* category, it clearly shows that 'meeting' and 'daily stand-up', mentioned in 28.8% and 16.5% of the 1450 descriptions respectively, are what the students perform most often. Moreover, meetings are more often with their 'supervisor' (10.3%) than with teachers (2.3%) or colleagues (2.2%). 'Test' is mentioned in 64.8% of all 250 descriptions of the *Testing* category and 'research' is mentioned in 37.8% of all 1810 descriptions of the *Research* category. 'Project plan', 'final report', 'reflection report', and 'personal development plan', are mentioned in 30.7%, 15.8%, 10%, and 7.3% of 2200 descriptions in *Academic Documentation* respectively, clearly showing what the university requires the students to write.

4.6 Technologies

The second text analysis performed on the descriptions tries to determine whether the students are using relevant and modern technologies in their internship. As mentioned in Section 3.3.2, we ran the descriptions of the relevant categories of each cohort through the text analysis pipeline (see Figure 3), and subsequently checked for each technology found, how many distinct students

Figure 11: Frequency of keywords in the category *Meeting*.

Figure 12: Frequency of keywords in the category *Testing*.

Figure 13: Frequency of keywords in the category *Research*.

Figure 14: Frequency of keywords in the category *Academic Documentation*.

² Although the keyword 'realize' occurred in 15.1% of 2140 descriptions, we cannot definitively claim that students do *realize*, since the Dutch word from which it derives ('maken')

can also be used as an auxiliary verb. The n-grams did not show any noteworthy patterns.

mentioned that particular technology. The results of this analysis for the SN and SE cohorts are shown in Figure 15 and Figure 16, respectively. The analysis of the descriptions of the BI students did not lead to any conclusive technologies (top mentioned technology was ‘Sharepoint’ by four distinct students).

As shown in Figure 15, several software engineering technologies are mentioned frequently by different SE students; ‘API’ is mentioned by 48.3% of students, just as ‘Database’ 48.3%, ‘Git*’ (meaning anything starting with git: ‘git’, ‘github’, ‘gitlab’, etc.) 41.4%, ‘REST’ and ‘RESTful’ by 41.4%, ‘Angular’ by 34.5%, ‘Jira’ by 31%, ‘CSS’ by 31%, and ‘HTTP’ and ‘HTTPS’ by 27.6%. Clearly, most of the SE students are indeed programming, even though our previous analysis was inconclusive on that matter. Moreover, most of them appear to be working on front-end (reactive webpages) and back-end (databases and api’s) applications. The students mention the programming languages used less; ‘JavaScript’ is mentioned by six distinct students (20.7% of total); ‘Java’, ‘Python’ and ‘C#’ are mentioned each by two distinct students (6.9% of total).

Figure 16 clearly shows that the SN students are mentioning ‘network’ rather often (84.6% of total), again showing that they appear to be working on what is expected. Other frequently mentioned technologies are: ‘server’ by 53.8%, ‘router’ by 46.2%, and ‘Windows’, ‘Cisco’, ‘Firewall’, and ‘Cluster’ by 30.8% each.

5. CONCLUSIONS AND DISCUSSION

5.1 Conclusions

The goal of this study was to gain cohort-based insight into which tasks computing students perform, and which technologies they use, while learning in the workplace. To this aim, we gathered and analyzed workplace activity data of 54 students over the course of their third-year semester-long internships in the computing industry. We performed descriptive analyses to gain insight into which categories of activities students performed most, which were programming, documentation and research.

Students added their own categories to the application to label activities. Most notably, they created categories for orientation and introduction activities, getting acquainted with a company efficiently and setting up their technical working environment. In our university, we do not prepare our students for these activities, e.g. by creating awareness on the importance of these type of activities when starting an internship. Additionally, we can use the results of this study for expectation management towards our students. When providing prospective interns with information and instruction, we could include data on which activities are performed most.

From the analysis of the difficulty of categories we can conclude that Academic documentation, Research, Implementation/configuration and IT Documentation are perceived as most difficult. These new insights could help us refine and improve computing curricula, especially the curriculum parts before the internship, e.g. by providing more opportunities for students to develop their knowledge and skills that are related to these difficult categories.

Subsequent text analysis gave us insight into students’ perceptions of the categories used to label activities. The user-generated descriptions of activities in the categories testing, research, meetings, and academic documentation are congruous with their labels, while the analysis of the descriptions of implementation/configuration, programming and IT documentation did not give conclusive results. Finally, we analyzed which technologies were used most by these students, which did not yield any conclusive

Figure 15: Technologies frequently mentioned by SE students.

Figure 16: Technologies frequently mentioned by SN students.

insights for the BI cohort, but we did find technologies for the SE and SN cohorts which were in line with what we expected based on the learning outcomes of these programs. Repeating this analysis every year has the potential to reveal technology trends which could be used to update the curriculum accordingly.

The content analysis of the user-generated descriptions, combined with the analyses of the most occurring categories, implies that the common joke stating that interns are mainly busy getting coffee for others is definitely not true for this group of computing students.

5.2 Limitations

Based on the results, we conclude it is feasible to use user-generated data to get insights into workplace activities of computing interns. However, the quality of this user-generated data creates the need to pair our conclusions with significant disclaimers. For example, if all students working with Java never explicitly wrote this down when registering their activities, it might seem Java is not used by a single student.

Registering time sheets at work is usually perceived as boring and bothersome. To some extent, we recognize ‘resistance’ to the workload involved in three main strategies students use to reduce the time needed to register their activities: 1) register work in very large portions (one or two activities per day), 2) creating custom ‘container categories’ to label activities, and 3) label activities as Easy to avoid a reflection prompt. All three strategies degrade the quality of the data and thus also the results obtained after analysis.

The text analysis we performed also has limitations. For example, in the category Programming, the variation of keywords used was very large. The text analysis pipeline we used did not allow to easily count and cluster different wordings for the same activities. Additionally, students do not always mention the technologies they used explicitly in the descriptions of their activities. Adjustments to the application (e.g., providing support questions in the input field for the descriptions or adding a field for technologies) could give higher quality data. A more advanced analysis of the user-generated texts can subsequently give more insightful results.

Finally, the students in this study opted in to participate, which can induce self-selection bias in the data. We analyzed data of merely 30% of all students performing an industry internship. We do not know whether their data was representative for the entire population. Furthermore, we only included students from one university in this study. All the above limitations have an influence on the obtained results and the generalizability of our conclusions.

5.3 Future Work

We plan to extend this study in the near future by recruiting a larger set of students, also by cooperating with other higher educational institutes. It would also be interesting to perform this type of study for other educational domains, such as economics, health or education. Furthermore, it would be interesting to include other, not user-generated, workplace data such as log files, etc. This was not feasible since the 54 students performed their internship at 54 different organizations. Moreover, relating the technologies used to the (type of) organization at which the internship was performed, could be an interesting analysis for the future.

Finally, the text analysis of the descriptions of activities could be extended by comparing the results with the results of a similar analysis of school-based descriptions (syllabi), e.g. to compare which technologies are used and the relative weight of categories in internships compared to the relative weight in the curriculum in order to identify ‘weak spots’ in the curriculum that differ most from what industry requires from computing graduates.

REFERENCES

- [1] 1st International Conference on Learning Analytics and Knowledge. 2011. Retrieved from <https://tekri.athabasca.ca/analytics/>
- [2] Jane Andrews and Helen Higson. 2008. Graduate employability, ‘soft skills’ versus ‘hard’ business knowledge: A European study. *Higher Education in Europe* 33, 4 (2008), 411–422. DOI: <https://doi.org/10.1080/03797720802522627>
- [3] Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Computational Linguistics in the Netherlands 2006: Selected papers from the seventeenth CLIN Meeting*. Vol. 7. LOT, Utrecht, 191–206.
- [4] Shane Dawson, Srecko Joksimovic, Oleksandra Poquet, and George Siemens. 2019. Increasing the impact of learning analytics. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19)*. ACM Press, New York, NY, USA, 446–455. DOI: <https://doi.org/10.1145/3303772.3303784>
- [5] Robert L. Dunbar, Molly J. Dingel, Xavier Prat-Resina, and Xavier Prat-Resina. 2014. Connecting analytics and curriculum design: process and outcomes of building a tool to browse data relevant to course Designers. *Journal of Learning Analytics* 1, 3 (Aug. 2014), 223–243. DOI: <https://doi.org/10.18608/jla.2014.13.26>
- [6] EC. 2016. ICT for work: Digital skills in the workplace.
- [7] EPRS | European Parliamentary Research Service. 2017. Digital skills in the EU labour market. *Technical Report*. European Parliament. DOI: <https://doi.org/10.2861/451320>
- [8] Maarten van Gompel and Antal van den Bosch. 2016. Efficient n-gram, skipgram and flexgram modelling with Colibri Core. *Journal of Open Research Software* 4, 1 (2016). DOI: <https://doi.org/10.5334/jors.105>
- [9] Julio Hernández-March, Mónica Martín del Peso, and Santiago Leguey. 2009. Graduates’ skills and higher education: the employers’ perspective. *Tertiary Education and Management* 15, 1 (Mar. 2009), 1–16. DOI: <https://doi.org/10.1080/13583880802699978>
- [10] Tobias Hüsing, Werner B. Korte, and Eriona Dashja. 2015. e-Skills in Europe. *Technical Report*. Retrieved from http://eskills-lead.eu/fileadmin/LEAD/Working_Paper_-_Supply_demand_forecast_2015_a.pdf
- [11] Mona Itani and Issam Srour. 2016. Engineering students’ perceptions of soft skills, industry expectations, and career aspirations. *Journal of Professional Issues in Engineering Education and Practice* 142, 1 (Jan. 2016). DOI: [https://doi.org/10.1061/\(ASCE\)EI.1943-5541.0000247](https://doi.org/10.1061/(ASCE)EI.1943-5541.0000247)
- [12] Adrian Kirkwood and Linda Price. 2014. Technology-enhanced learning and teaching in higher education: what is ‘enhanced’ and how do we know? A critical literature review. *Learning, Media and Technology* 39, 1 (Jan. 2014), 6–36. DOI: <https://www.tandfonline.com/doi/full/10.1080/17439884.2013.770404>
- [13] Fabienne M. van der Kleij, Jorine A. Vermeulen, Kim Schildkamp, and Theo J.H.M. Eggen. 2015. Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice* 22, 3 (Jul. 2015), 324–343. DOI: <http://www.tandfonline.com/doi/full/10.1080/0969594X.2014.999024>
- [14] Tobias Ley, Ralf Klamma, Stefanie Lindstaedt, and Fridolin Wild. 2016. Learning analytics for workplace and professional learning. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*. ACM Press, New York, NY, USA, 484–485. DOI: <https://doi.org/10.1145/2883851.2883860>
- [15] Gonzalo Méndez, Xavier Ochoa, and Katherine Chiliza. 2014. Techniques for data-driven curriculum analysis. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge - LAK '14*. ACM Press, New York, NY, USA, 148–157. DOI: <https://doi.org/10.1145/2567574.2567591>
- [16] Wim J. Nijhof, Loek F.M. Nieuwenhuis (Eds.). 2008. The learning potential of the workplace. Sense Publishers.
- [17] Adolfo Ruiz-Calleja and Luis Pablo Prieto. 2017. Learning analytics for professional and workplace learning: a literature Review. (Sep. 2017). DOI: <https://doi.org/10.1007/978-3-319-66610-5>
- [18] Kim Schildkamp and Wilma Kuiper. 2010. Data-informed curriculum reform: which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education* 26, 3 (2010), 482–496.
- [19] Andreas Schmidt and Christine Kunzmann. 2006. Towards a human resource development ontology for combining competence management and technology-enhanced workplace learning. In *OTM 2006, LNCS 4278*. Springer, Berlin, Heidelberg, 1078–1087. DOI: https://doi.org/10.1007/11915072_10
- [20] Melody Siadaty, Jelena Jovanović, Dragan Gašević, Nikola Milikić, Zoran Jeremić, Liaqat Ali, Aleksandar Giljanović, and Marek Hatala. 2012. Semantic web and linked learning to support workplace learning. In *CEUR Workshop Proceedings*, Vol. 840.

- [21] Esther van der Stappen. 2018. Workplace learning analytics in higher engineering education. In *2018 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 15–20. DOI: <https://doi.org/10.1109/EDUCON.2018.8363102>
- [22] Esther van der Stappen and Ilya Zitter. 2017. Design propositions for technology-enhanced workplace learning. In *Proceedings of EAPRIL 2017*. Hämeenlinna, Finland, 37–51.
- [23] Lisette Toetenel and Bart Rienties. 2016. Analysing 157 learning designs using learning analytic approaches as a means to evaluate the impact of pedagogical decision making. *British Journal of Educational Technology* 47, 5 (Sep. 2016), 981–992. DOI: <https://doi.org/10.1111/bjet.12423>
- [24] Päivi Tynjälä. 2013. Toward a 3-P Model of Workplace Learning: A Literature Review. In *Vocations and Learning*, 6, 1 (Apr. 2013), 11–36. DOI: <https://doi.org/10.1007/s12186-012-9091-z>
- [25] Päivi Tynjälä, Virpi Slotte, Juha Nieminen, Kirsti Lonka, and Erkki Olkinuora. 2006. From university to working life: graduates' workplace skills in practice. *Higher education and working life: collaborations, confrontations and challenges*. (Jan. 2006), 73–88.
- [26] Technology enhanced learning | Higher Education Academy. 2018. Retrieved from <https://www.heacademy.ac.uk/individuals/strategic-priorities/technology-enhanced-learning>
- [27] Vathsala Wickramasinghe and Lasantha Perera. 2010. Graduates', university lecturers' and employers' perceptions towards employability skills. *Education + Training* 52, 3 (Apr. 2010), 226–244. DOI: <https://doi.org/10.1108/00400911011037355>
- [28] Dylan Wiliam. 2011. What is assessment for learning? *Studies in Educational Evaluation* 37, 1 (Mar. 2011), 3–14. DOI: <https://doi.org/10.1016/J.STUEDUC.2011.03.001>
- [29] Ilya Zitter, Aimée Hoeve, and Elly de Bruijn. 2016. A design perspective on the school-work boundary: a hybrid curriculum model. *Vocations and Learning* 9, 1 (Feb. 2016), 111–131. DOI: <https://doi.org/10.1007/s12186-016-9150-y>